ANALYSIS OF THE AIRBNB SCENE IN BARCELONA

Course: Multivariate Analysis
Dataset: Listings.csv (07 June 2022) for Barcelona from insideairbnb.com
Team members: Dorian Bauschke, Pau Comas, Marçal Garcia, Liam Glennie, Clara Molins and Catalina Priescu (group 6)
Partial Delivery: 24/10/2022

INDEX

1. Motivation of the work and general description

Our main objective is to analyse the current situation of Airbnb listings in Barcelona and surrounding areas, i.e. the metropolitan area of Barcelona. Our aim is to understand what characteristics of a listing plays a role in its price and rating. Our main target variable is price, however we will be interested in potentially using rating as our secondary target.

By conducting this analysis we seek to demonstrate several/a number of initial hypotheses. On the one hand, we believe that for price there could be certain features that have a greater influence on it than others (such as: number of bedrooms, beds, location, number of reviews…etc.) This assumption needs to be tested and verified. On the other hand, we are also interested in identifying which features have a noticeable impact on the average reviews per month for each listing as well as their availability. Using this data, we might theoretically anticipate the success of future Airbnb listings.

A secondary goal of this project is to learn how to manipulate data by applying preprocessing techniques to deal with outliers and missing values, univariate and multivariate analysis. To further comprehend our data, we will make use of statistical tools, such as PCA, clustering, MCA, MFA and association rules among others. To carry out this project, we will use the programming language R and the IDE R studio, commonly used for statistical problems.

2. One Pager

In this project, we aim to analyse the Airbnb scene of Barcelona and surrounding areas. Our main goal is to identify the most contributing factors to the price of a listing. Our dataset has 16646 listings and 74 attributes containing information about the host, physical features of a listing, booking process and price.

First, we removed many variables that provided us with irrelevant information, like URLs or details about the scrape used to obtain the data. Next, we modified the variables by creating groups of modalities, applying format rules to dates, removing troublesome characters and one-hot encoding amenities into a reduced set of variables.

Our univariate analysis points out a substantial number of NAs in our dataset. We implemented different manners of imputation. Some were done manually by checking the website; others involved slightly more sophisticated methods, such as MICE or imputing with a normal distribution.

The bivariate analysis leads us to our first exciting findings. We see that the review score variables do not seem to have much influence on the price of a listing. However, the neighbourhood does, a trend that continues throughout our analysis.

The factorial methods provide us with the most attractive insights yet. In PCA, we observe how the number of guests a listing accommodates impacts its price. We also find that Eixample has the highest median price as it is based in the centre of Barcelona and the average listing accommodates four people. PCA also allows us to distinguish between short and long-term listings. However, our PCA findings are inconclusive, and the explained variance in the first two dimensions is around 35%.

MCA highlights the distinction between luxury and basic listings. Luxury listings are deemed to be listings with pools, barbecues, and private parking located in wealthier neighbourhoods. Whereas basic listings may lack basic amenities such as WiFi and TV, have shared bathrooms and are likely to be in areas further from the city centre.

To be able to perform MFA, we create groups of variables depending on the type of data and the meaning it has. We find that out of the groups, both numerical and categorical host information, together with a listing's categorical physical aspects, explain the most variance in our dataset.

In this partial part of the project, we finish with association rules. This method confirms once more that the neighbourhood plays a noteworthy role in a listing's price and rating. More expensive listings in central areas tend to receive higher reviews. Furthermore, it seems that, in general, users are happy with what they pay for, not creating any false expectations of finding an incredible listing at a low price.

3. Data source presentation

The database was acquired from the URL http://insideairbnb.com/get-the-data/. The dataset downloaded was 'listings.csv.gz' from Barcelona, data produced on 07 June 2022.

Date Compiled	Country/City	File Name	Description
07 June, 2022	Barcelona	listings.csv.gz	Detailed Listings data
07 June, 2022	Barcelona	calendar.csv.gz	Detailed Calendar Data
07 June, 2022	Barcelona	reviews.csv.gz	Detailed Review Data
07 June, 2022	Barcelona	listings.csv	Summary information and metrics for listings in Barcelona (good for visualisations).
07 June, 2022	Barcelona	reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
N/A	Barcelona	neighbourhoods.csv	Neighbourhood list for geo filter. Sourced from city or open source GIS files.

Ref 1: Screenshot of the page the data was extracted from.

The original dataset contains 16649 tuples and has a total number of 74 variables. Despite this, at the very beginning we decided in agreement with the professors to do a first pruning leaving out some features that seemed blatantly irrelevant, such as URL strings of listings, images or numerical/categorical attributes with almost each value as missing data.

Numeric Variables:

1. Host Response Rate (*host_response_rate*)
2. Host Acceptance Rate (*host_acceptance_rate*)
3. Host Total Listing Count (*host_total_listings_count*)
4. Price (*price*)
5. Minimum / Maximum Nights (*minimum_nights/ maximum_nights*)
6. Accommodates (*accommodates*)
7. Bathrooms (*bathrooms*)
8. Bedrooms (*bedrooms*)
9. Beds (*beds*)
10. Latitude (*latitude*)
11. Longitude (*longitude*)
12. Availability 30 (*availability_30*)
13. Availability 365 (*availability_365*)
14. Number of reviews (*number_of_reviews*)
15. Number of reviews last twelve months (*number_of_reviews_ltm*)
16. Review Scores Accuracy (*review_scores_accuracy*)
17. Review Scores Cleanliness (*review_scores_cleanliness*)
18. Review Scores Checkin (*review_scores_checkin*)

19. Review Scores Communication (*review_scores_communication*)
20. Review Scores Location (*review_scores_location*)
21. Review Scores Value (*review_scores_value*)
22. Reviews per Month (*reviews_per_month*)
23. Calculated Host Listings Count (*calculated_host_listings_count*)
24. Calculated Host Listings Count Entire Homes (*calculated_host_listings_count_entire_homes*)
25. Calculated Host Listings Count Private Rooms (*calculated_host_listings_count_private_rooms*)
26. Calculated Host Listings Count Shared Rooms (*calculated_host_listings_count_shared_rooms*)

Binary Variables:

1. Host has Profile Pic (*host_has_profile_pic*)
2. Host Identity Verified (*host_identity_verified*)
3. Host is Superhost (*host_is_superhost*)
4. Instant Booking (*instant_bookable*)
5. Has Availability (has_availability)

Qualitative Variables:

1. Host Since (*host_since*)
2. Host Location (*host_location*)
3. Host Response Time (*host_response_time*)
4. Host Verifications (*host_identity_verified*)
5. Neighborhood Group Cleansed (*neighbourhood_group_cleansed*)
6. Property Type (*property_type*)
7. Room Type (*room_type*)
8. License Type (*license*).
9. Review Scores Rating (*review_scores_rating*)
10. Amenities (amenities)
11. First Review (first_review)
12. Last Review (last_review)
13. Host verifications (host_verifications)

4. Data structure and metadata

4.1. Meaning of the rows

The information outlines the features of Airbnb listings located in and around Barcelona alongside with details regarding their respective hosts. Each listing has been assigned with a unique identifier and is further focused on mainly the following areas: information concerning the tenant (around 18 variables describing the tenant) and the property they offer (13 outlining property features), their prices (which is the main target feature in our analysis) and reviews (18 outlining review specifications).

Note: The original database has a number of 16649 entries. We have excluded some cases of extreme outliers in parts of our study because they were harmful to our analysis.

4.2. Metadata table

The stipulated metadata table comprises 48 variables that encompass information about the data recorded in the original database. There is an assigned brief meaning to each of the individuals as well as a data type, measuring unit, missing code, measuring procedure, range and role method, if applicable. In terms of modalities, it has been recorded further where the data allowed us to do so (i.e. it was qualitative). The majority of the data was preserved exactly as received from the source, but where necessary, we performed a number of changes by classifying qualitative in a more clear and succinct manner. Finally, we added a column that describes the imputation technique for each variable.

Metadata table:
https://docs.google.com/spreadsheets/d/17e6PucvXvM_K029dXyKXRXwXO4JrG5FCzGeN9dh0kYo/edit?usp=sharing

4.3. Inclusion and exclusion criteria for rows and columns

As we mentioned above, the overall aim of this paper is to investigate and test several dependencies among variables in further detail. Primarily, we wanted to focus on the impact of variables that characterize Airbnb per se, on its price, availability and the number of reviews they receive each month. Consequently, we chose the data in such a manner that it provides us with the information we require to carry out these research and analysis. The following variables were removed from the dataset for being redundant or irrelevant:
id, listing_url, scrape_id, last_scraped, name, description, neighborhood_overview, picture_url, host_id, host_url, host_name, host_about, host_thumbnail_url, host_picture_url, host_neighbourhood, host_listings_count, host_verifications, host_has_profile_pic, host_identity_verified, neighbourhood, neighbourhood_cleansed, minimum_minimum_nights, maximum_minimum_nights, minimum_maximum_nights, maximum_maximum_nights,

minimum_nights_avg_ntm maximum_nights_avg_ntm, calendar_updated, availability_60, availability_90 and *calendar_last_scraped.*

These variables were removed using Excel and were saved in the document "AIRBNB_Variables_Used_D2.xlsx". The rest of modifications of the dataset were performed with R.

5. Preprocessing

We divided the preprocessing process into two different steps: modification of variables and imputation of missing values.

5.1. Modification of variables

The first step was to modify some of the attributes of the original dataset in order to solve character problems (ex: issues with accents), simplify some of the attributes by creating new categories which grouped listings in a better way and finally one-hot encode some variables which contained a list of 'subattributes' with the aim of smoothing later analysis.

The modified variables and their respective changes were the following:

- **Host since**: We decided to only keep the year instead of day, month and year information because we thought it was unnecessary for our analysis, which is mainly focused on the price of the listings.
- **Amenities**: Each listing had a list of amenities. The total number of distinct amenities was more than 1500. Knowing this, what we did was choose some of the most common and crucial ones from all the listings and created boolean categorical variables for each of them. In other words, we did a one-hot encoding. Particularly, the chosen amenities kept for further study were the following:
    - Wifi (*wifi*)
    - Long term stays (*longTermStays*)
    - Air conditioning (*aircon*)
    - Heating (*heating*)
    - Tv (*tv*)
    - Hair dryer (*hairDryer*)
    - Host greets you (*hostGreets*)
    - Patio or balcony or backyard (*outdoorSpace*)
    - Parking on premises (*parkingOnPremise*)
    - Pool (*pool*)
    - Barbecue (*bbq*)
- **Host verification**: This variable also included a list for each listing. Each host can choose to let the verification be made through phone, email or work email. What we did, similarly to the amenities case, was convert this variable into three by creating 3 boolean categorical variables called:
    - *phone_verification*
    - *email_verificaiton*
    - *work_email_verification*
- **Neighbourhood group cleansed**: We removed all accents from the names of the different districts of Barcelona because R did not read them correctly.

- **Property type**: We grouped the different property types into just three to smooth later analysis, following the conclusions of the barplots obtained previously, during univariate analysis:
    - Entire rental unit
    - Private room in rental unit
    - Others
- **Bath type**: Originally, we had a feature called bathroom text that contained the type and number of bathrooms per listing. We divided the types into 3 categories:
    - Private
    - Shared
    - Unknown
- **Num baths**: Then, we extracted the number of bathrooms from bathroom text to create a numerical variable. Due to problems with R, bathrooms with an unknown number of bathrooms (i.e NA's), are codified as -1. These instances will be imputed in a future step.
- **License**: Every host has a different license. Depending on its first characters the license can be grouped into different classes:
    - Hotel
    - Independent
    - Exempt
    - Unknown
    - Other
- **Host Location**: The different values for this attribute were a mix of countries and cities from all around the world and we decided to group them by countries, specifically we grouped them as:
    - ES (Spain)
    - FR (France)
    - IT (Italy)
    - GB (Great Britain)
    - US (United States)
    - Others
- **First review**: We decided to keep the year and month instead of day, month and year because we thought the day information was unnecessary and too sparse for our analysis operations.
- **Last review**: Exactly the same as the procedure we did on variable First Review.
- **Price**: We removed the dollar sign and converted the variable into a number for obvious practical reasons.

Additionally, there were some variables that were categorical with two modalities: "t" and "f". These attributes were changed to boolean variables converting the modalities to True and False.

5.2. Missing Values

From the univariate analysis, which will be discussed later, we detected that our dataset had a large number of NAs in many variables. Therefore, the method we chose to impute the NAs differs for each attribute. For some of them, particularly variables which had a little amount of NAs, we checked the information on the Airbnb webpage. For other variables which had a larger number of missing values, we followed some of the imputation methods seen in class such as: imputing as unknown, imputing the mean, using the MICE method and others. The following list contains all variables that had NAs and the imputation method chosen to deal with them:

- **Host since**: Check Airbnb webpage
- **Host location**: Replace with ES (Spain)
- **Host response rate**: Impute using truncated random normal distribution
- **Host acceptance rate**: Impute using truncated random normal distribution
- **Host is superhost**: Check Airbnb webpage
- **Host total listings count**: Check Airbnb webpage
- **Host has profile picture**: Check Airbnb webpage
- **Host identity verified**: Check Airbnb webpage
- **Number of baths**: Impute NA with MICE using beds and bedrooms
- **Bedrooms**: Impute NA with MICE using beds and num_baths
- **Beds**: Impute NA with MICE using bedrooms and num_baths
- **First review**: Impute NA to 'Unreviewed'
- **Last review**: Impute NA to 'Unreviewed'
- **Review scores rating**: Impute with the mean of the rest of review scores
- **Review scores accuracy**: Impute with the mean or using truncated random normal distribution when the listing had at least one review
- **Review scores cleanliness**: Impute with the mean or using truncated random normal distribution when the listing had at least one review
- **Review scores checkin**: Impute with the mean or using truncated random normal distribution when the listing had at least one review
- **Review scores communication**: Impute with the mean or using truncated random normal distribution when the listing had at least one review
- **Review scores location**: Impute with the mean or using truncated random normal distribution when the listing had at least one review
- **Review scores value**: Impute with the mean or using truncated random normal distribution when the listing had at least one review
- **License**: Impute NA as unknown
- **Reviews per month**: Impute NA to 0

6. Univariate analysis

In terms of univariate analysis, we divided the variables into two: numerical variables and categorical variables, and performed the same analysis on each of the two. For categorical variables, we decided to do the univariate analysis before any preprocessing, in order to use univariate analysis as a support tool for preprocessing in the best way the categories if needed. Therefore, we believed that the Barplot that includes the "N/As" is the best form of the general evaluation of the data distribution in each category and being able to clearly see the proportion of NAs. For the numerical variables, the division went even further and we proposed a number of plots to be carried out:

- boxplot for outlier detection and analysis
- continuous numerical: kernel density plots
- discrete numerical: histograms
- number of missing data graph (from the slides)

Because of the high number of variables analyzed from the dataset, now we include only the most relevant univariate results and leave the rest of them in the Annexes.

Property type:

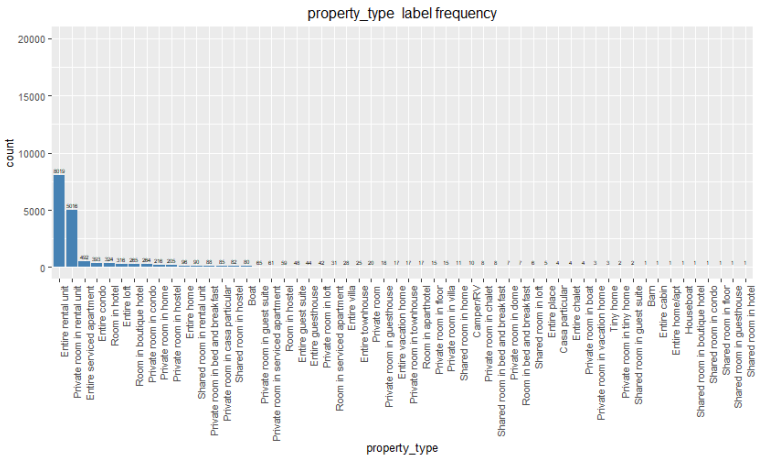


Image 1: Property Type Modalities

Roughly 50% of the values belong to the "Entire rental unit" modality. As we stated in the preprocessing section, we then limited the categories to "Entire rental unit", "Private room in rental unit" and "Others".

Price:

Prices range from 8 to 90000, but 95% of values are between 0 and 500. Below are two boxplots, one boxplot for the full range of prices and the second for until 10000. We can observe many outliers, not errors, because looking further into the data we can see that they are

specific luxury listings that have very high and exclusive prices. The problem with luxury listings is that now and when we do the bivariate analysis they ruin the plots' clarity.

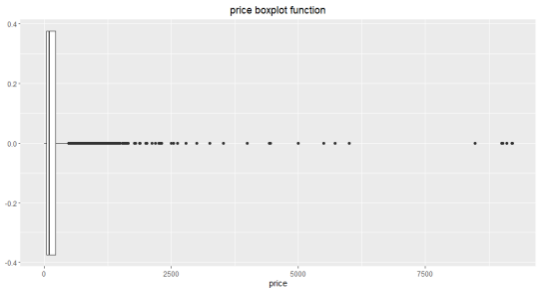


Image 2: Reduced Price Boxplot

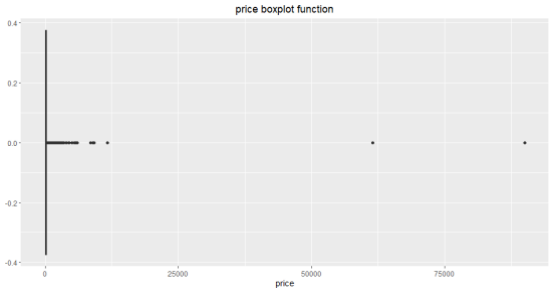


Image 3: Full Price Boxplot

Beds:

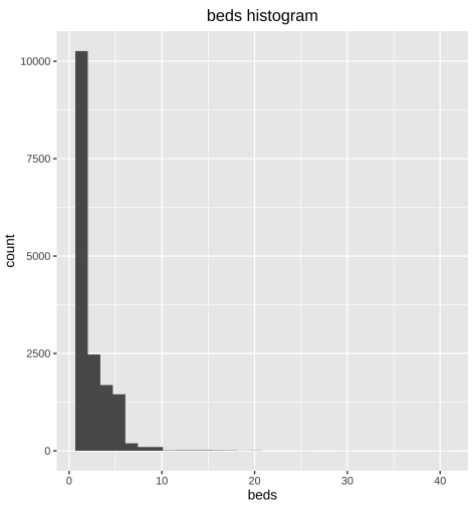


Image 4: Beds Histogram

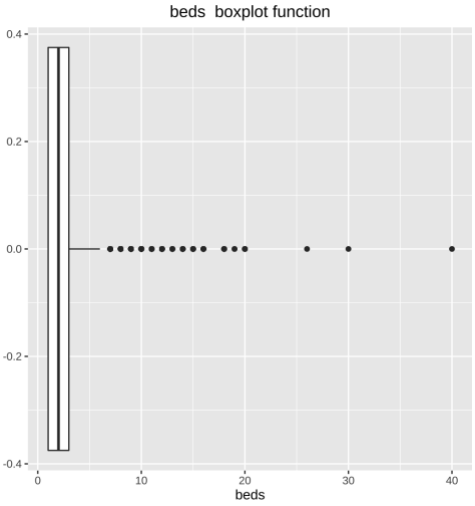


Image 5: Beds Boxplot

As we can see most of the listings have few beds, 89% between 1 and 4, but as seen in "price" and in other univariate descriptors of the listings, outliers appear as a consequence of keeping luxury and exclusive listings as part of the dataset.

Number of reviews:

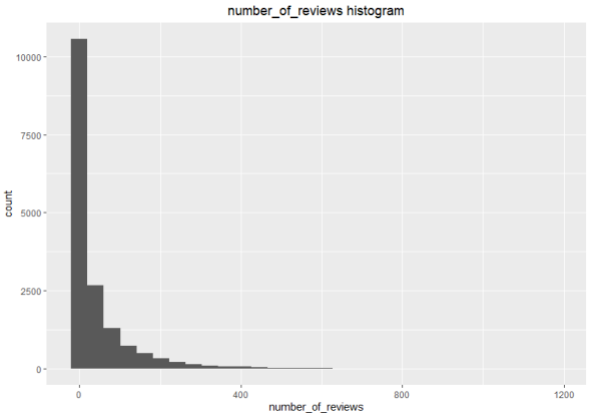


Image 6: Number of Reviews Histogram

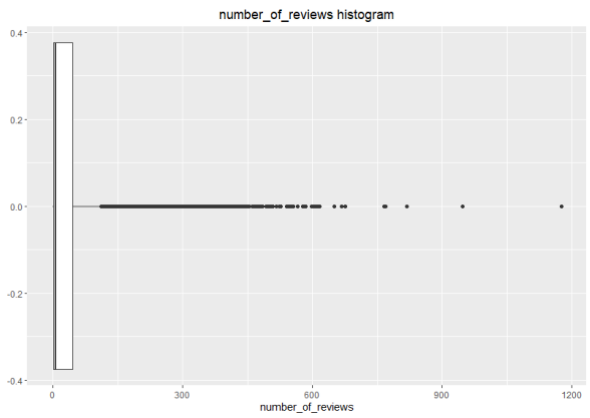


Image 7: Number of Reviews Boxplot

In number_of_reviews and derivatives in the dataset (per month, per year) there is a clear bias toward a low number of reviews. 3681 of 16646 listings have not been reviewed yet (have 0 reviews). Many outliers appear, belonging in almost every case to hostel rooms that were listed on Airbnb.

Host response time:

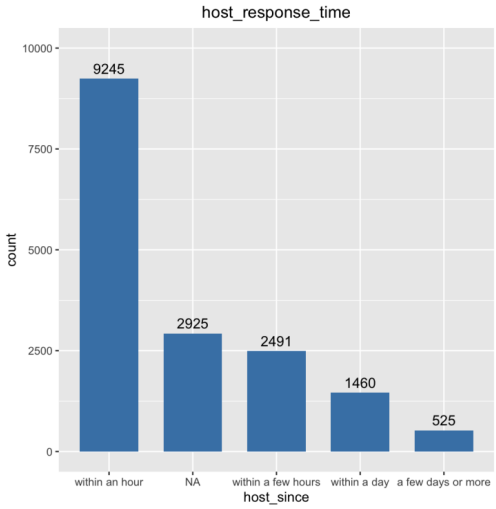


Image 8: Host Response Time Modalities

Most of the hosts reply within an hour. There were 2920 NAs that, as we mentioned, we imputed as "Unknown".

Review scores rating:

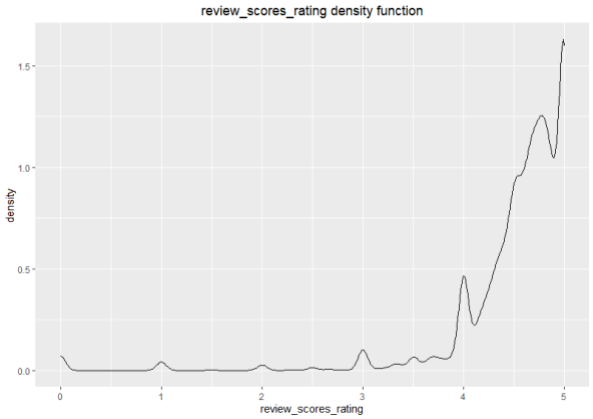


Image 9: Review Score Rating Histogram

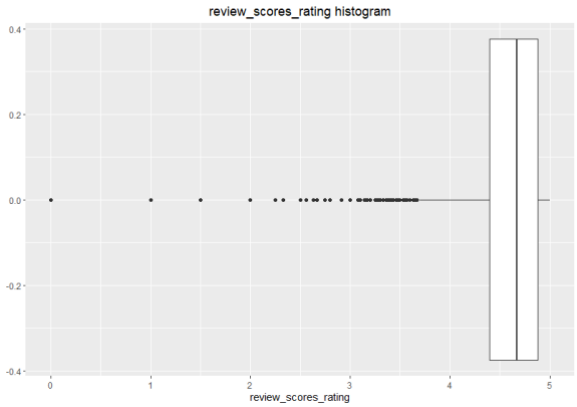


Image 10: Review Score Rating Boxplot

Most of the average review scores are closer to 5, outliers are for low reviews for specific listings. 113 original listings had an average of 0. Most of the 113 listings have had only 1 review, some of them 2 and much less 3+. The mean average rating is 4.64.

In other review metrics, such as cleanliness, location, or others, the pattern is very similar. Most of the listings have very high review scores in almost everything, with minority instances having low scores.

Neighbourhood group cleansed:

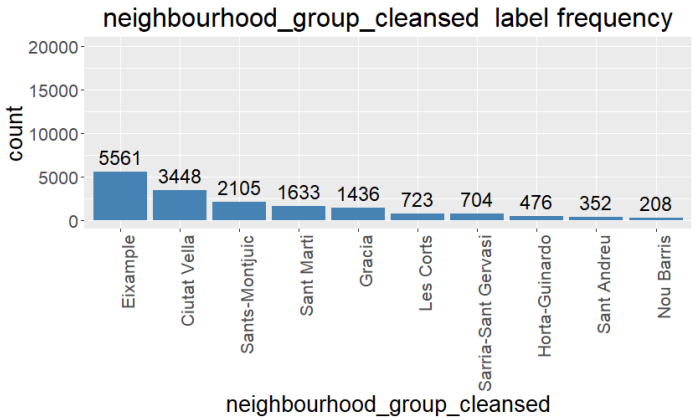


Image 11: Neighbourhood Modalities

Eixample is the neighborhood with the most listings, with 5561. This equals to 33% of the total listings. As we can see there are up to ten different neighborhoods. The most popular neighborhoods are logically the ones closer to the center and with more tourist sites.

Host is superhost:

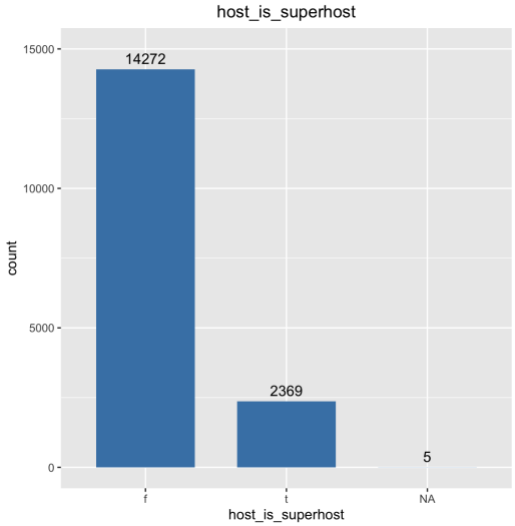


Image 12: Host is Superhost Barplot

This variable doesn't have much complication to analyze rather than the 5 NAs it had before imputation, but it shows exciting relations when applied to bivariate and other analyses later.

7. Bivariate analysis

Based on the insights obtained in the univariate analysis, we decided to find some patterns and relations of interest between variables in the bivariate analysis. We focused on relations with price and ratings attributes, even though we explored other alternatives as well. We highlight in this part the most relevant plots and we leave the rest the Annex..

Review scores rating and host is superhost:

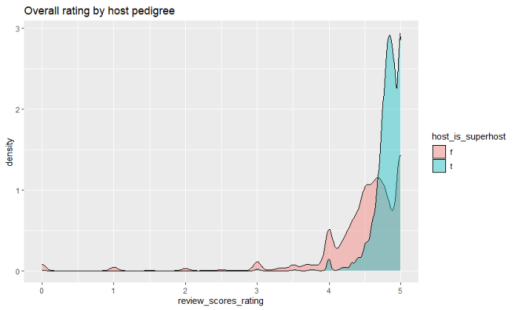


Image 13: Rating by host pedigree before imputation

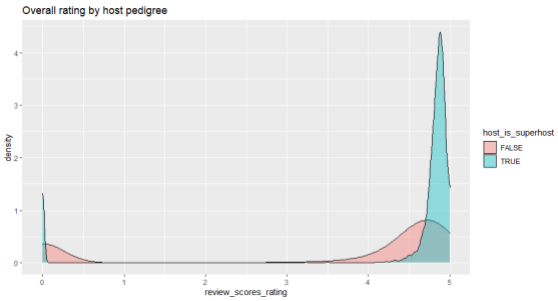


Image 14: Rating by host pedigree after imputation

Density plots before (left) and after (right) imputation. We can note that overall ratings, even though they are good almost always, are better if a host has the superhost badge.

Different review scores attributes correlation:

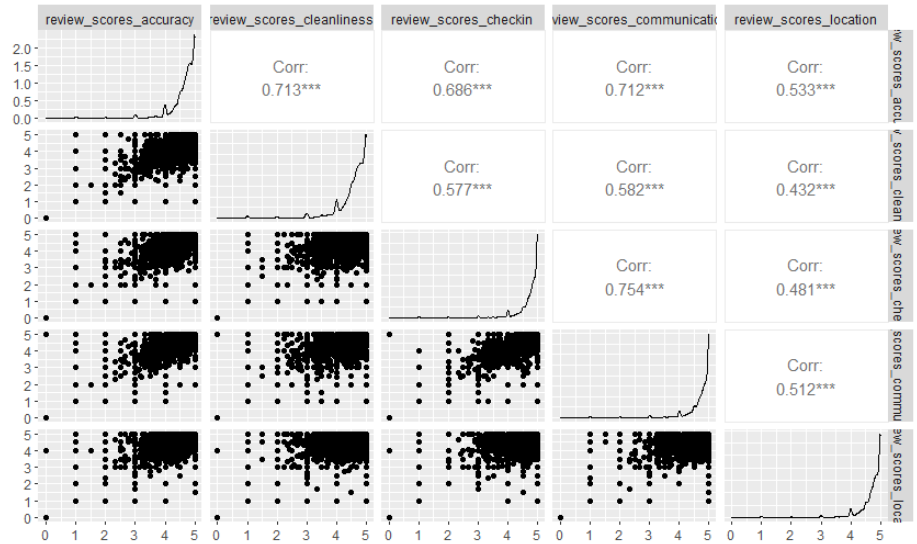


Image 15: Correlation between review scores

We thought that probably positive ratings of a listing could be correlated, even though they were about different things (location in the city, cleanliness…). We can see in the image above that we were right, being the check-in rating and the communication the most correlated (0,782) and location and cleanliness the less (0.482), although it is still significant.

Price and host is superhost:

Image 16: Price < 10000 by host pedigree Boxplot

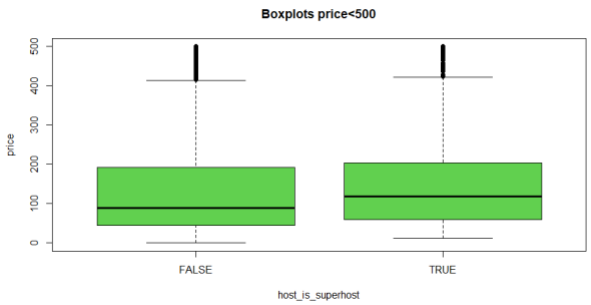


Image 17: Price < 500 by host pedigree Boxplot

The outliers preserved in price make it difficult to very generic plots around price. Nonetheless, filtering price to a maximum of 500 (right image) allows us to see things a bit clearer, even though the relation between price and being a superhost seems very slightly positive.

Price and review scores:

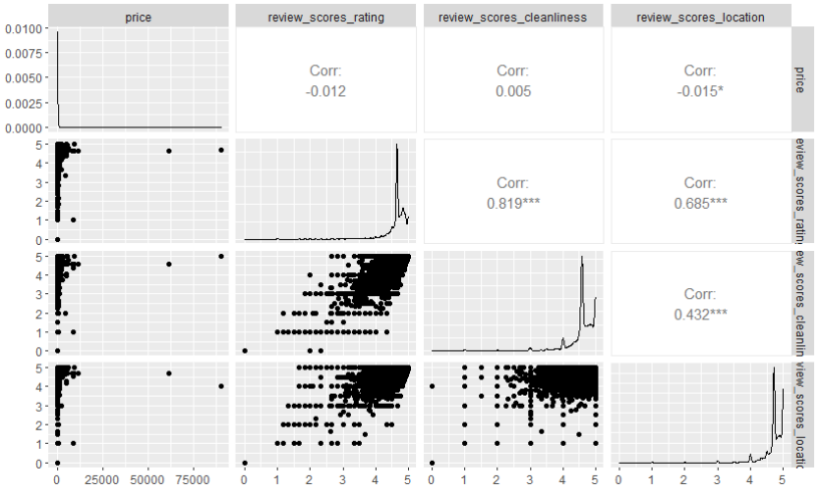


Image 18: Correlation between review scores and price

One of the surprises we found in many plots is that price is not correlated with the ratings. Seems that clients value the accommodations based on their price quality level experience.

Review scores rating and host response time:

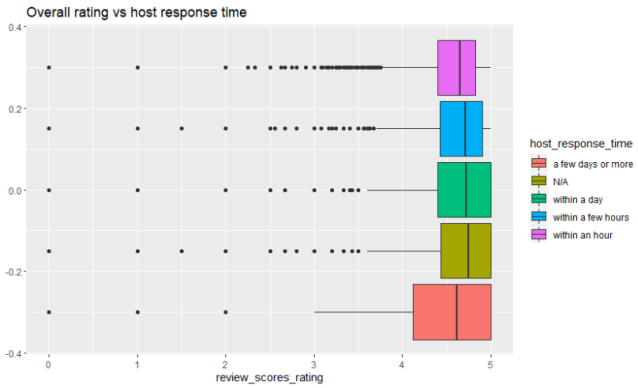


Image 19: Rating by Host Response Time Boxplot

As seen in the image above and other analyses, most of the ratings are good regardless of the classification of the listings, in this case divided by *host_response_time*. However, with this plot we could say that there is a tendency of the overall ratings of a listing to be a bit lower if the host takes "a few days or more" to reply.

Price and neighbourhood group cleansed:

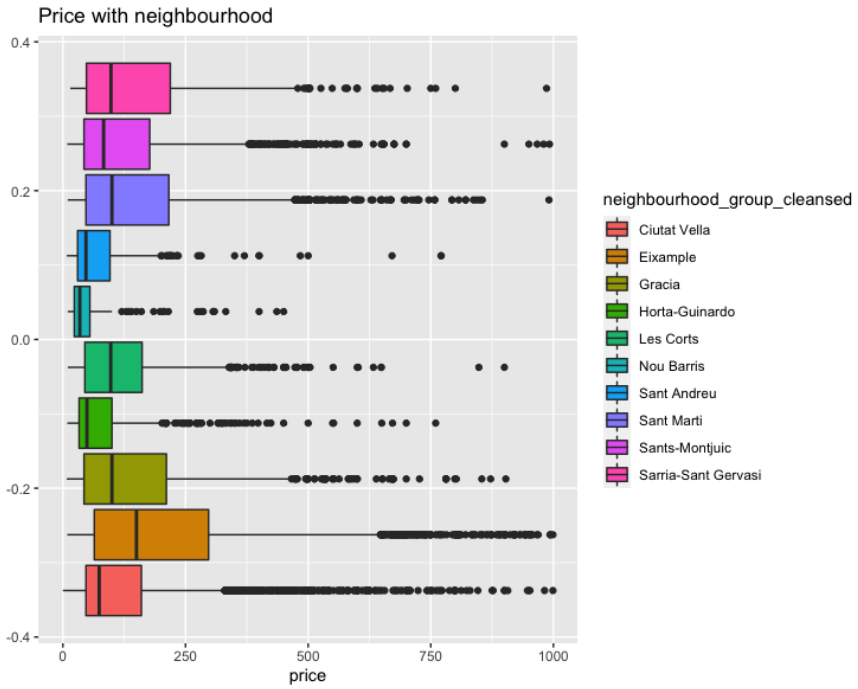


Image 20: Price by Neighborhood Boxplot

With the price filtered up to a maximum value of 1000, we can see some trends on the price boxplots based on the district of Barcelona the listings belong. One of the poorest districts in Barcelona is Nou Barris, and we see that the prices of the accommodations are a bit lower. We also see a lower tendency for Horta-Guinardo, and we believe that the reason could be the

location of the zone, far from the center of Barcelona. The most expensive zone seems to be the Eixample, centric and with relatively modern buildings.

Review scores location and neighbourhood group cleansed:

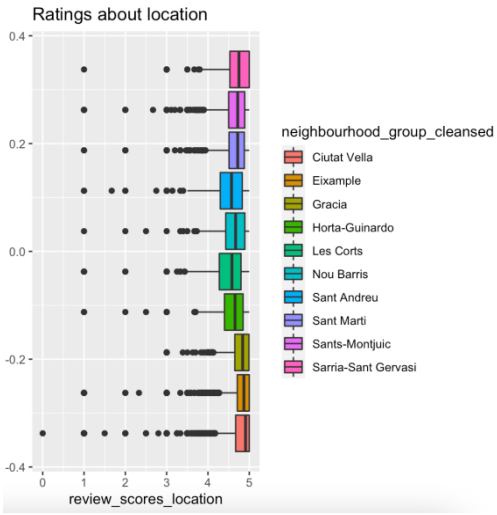


Image 21: Location Rating by Neighborhood

Ratings based on the location of the district are almost always very high, with some exceptions in outlier instances. The slight difference in boxplot results seems logical looking at the map of Barcelona. Nou Barris and Horta-Guinardo are districts far from the most popular touristic zones and Ciutat Vella and Eixample exactly the opposite.

8. Conclusions of univariate and bivariate analysis

After analysing our preprocessed dataset we can extract several conclusions about our listings data. Firstly, we have many categorical variables highly skewed to one category, such as *host_has_profile_pic*, *has_availability* and amenities like *tv*, *wifi* in a positive way of having them and *outdoorSpace* or *Pool* in a negative way of not having them. This is valuable information for us to take into account when looking at factorial methods or association rules results.

On the other hand, there are many attributes very similar between each other and that its distributions follow the same patterns. The different variables of *reviews_per_month* or *has_availability* show us that, where looking closer we could see that outliers correspond to the same special listings in many cases.

Finally, analysis around our targeted variables *price* and the different *review_scores* show us that, although the distributions of the variables are very narrow (*price* towards prices closer to 0 and *review_scores* to high reviews) and have some remarkable outliers we decided to keep, we have found many possible relationships and correlations between them and other variables.

For instance, location of the listing (*neighborhood_group_cleansed*) seems to have influence both in *price* and *review_scores*, especially in *review_scores_location*. Between *review_scores*, we've seen correlation between all the different metrics and host descriptors such as if the host is superhost or the response time of the host. One of the surprises is that looks like *price* doesn't have any relation with the *ratings* given by clients, strengthening our initial hypothesis that the listings and host characteristics may be important in order to predict the target variables by separate.

9. PCA analysis

Before discussing the results obtained by the PCA let us first explain the chosen numerical variables with which we decided to execute it.

The total number of numerical variables in our dataset was 28, which is a considerable amount and removing our target variable, price, (which is not included in the PCA) leaves us with 27. So in order to reduce the number of attributes and also improve the PCA's results, we checked the correlations between all numeric variables.

By doing so we saw, as expected, that *calculated_host_listings_count_entire_homes*, *calculated_host_listings_count_private_rooms*, *calculated_host_listings_count_ _shared_rooms* and *calculated_host_listings_count* were all highly correlated.
Note that the last of these variables is the sum of the other three.

In addition, we also noted that *review_scores_rating*, *review_scores_checkin*, *review_scores_accuracy*, *review_scores_cleanliness*, *review_scores_communication*, *review_scores_location* and *review_scores_value* were as well highly correlated.
Despite the first one being the average of the rest, we decided to keep *review_scores_value* because we believe that it is the most related to price. *Review_scores_value* measures how the listing is rated in terms of value for price.

Finally, we detected as well that *number_of_reviews_l30d*, *number_of_reviews_ltm* and *reviews_per_month* were highly correlated.

Therefore, taking into account the above stated, we decided to only keep the last attributes of all sets exposed (*calculated_host_listings_count*, *review_scores_value* and *reviews_per_month*) in addition to all the other numeric variables, leaving us with 16.

While performing the PCA analysis there were some variables that even when applied log() or sqrt() function ruined the plots and removed importance to the main cloud of points. For that reason, we decided to remove the following rows from the dataset used for the PCA and for the supplementary plots as well:

- Listings with a price higher than 5000
- Listings with more than 60 reviews per month

In the following pages the PCA is exposed in detail.

9.1. Scree plot and selection of principal components

After applying the PCA method to the numerical variables explained above, the first thing we checked was their individual contribution to the explanation of the total variance. The scree plot below shows the exact numbers obtained for the first 10 dimensions (the most significant ones).

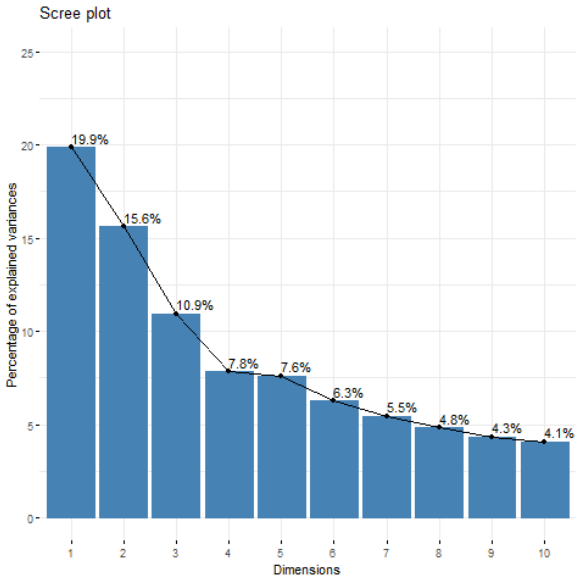


Image 22: PCA Scree Plot

Afterwards we looked for the number of principal components necessary to explain up to a 60% of the variance, and we decided to keep the first five principal components, which particularly explain 63% of the total variance. The following plot shows the cumulated contributions:

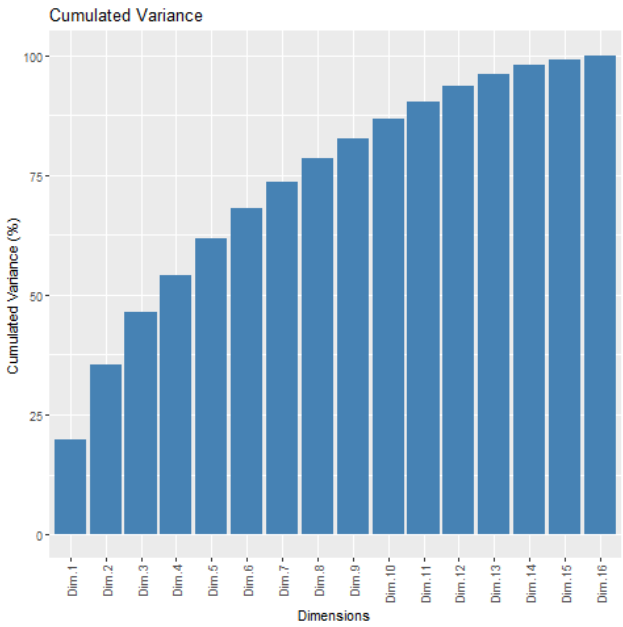


Image 23: PCA Cumulated variance plot

9.2. Factorial map visualization

In this section we will first show the variables factor map and afterwards we will expose the factor map for listings. The following plot shows graphically the relationship between each variable and the first two principal components, from now on we will call them PC1 and PC2.

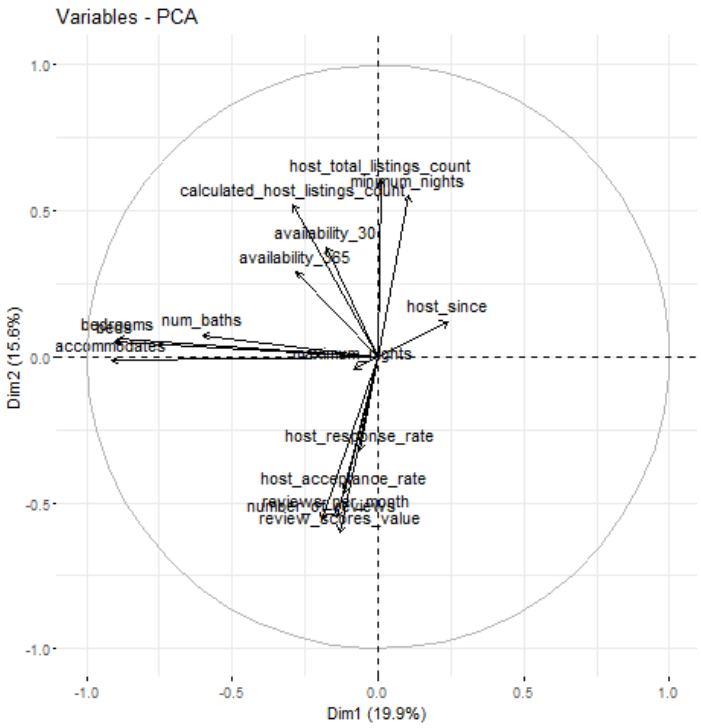


Image 24: PCA variables factor map

We can deduce that the variables *bedroom*, *accommodates*, *num_baths* and *beds*, which are all related to the number of people who can stay in each listing, are quite correlated and seem to have a high negative contribution to PC1.

Then, it can also be seen that those variables related to the quality of the host (reviews and host rates) have a high contribution to PC2 as well as *host_total_listings_count*, which is almost over the Dim2 line and *minimum_nights*.

The rest of the variables contribute a bit more to PC2 rather than PC1 but in a very little amount.

The following plots prove these statements:

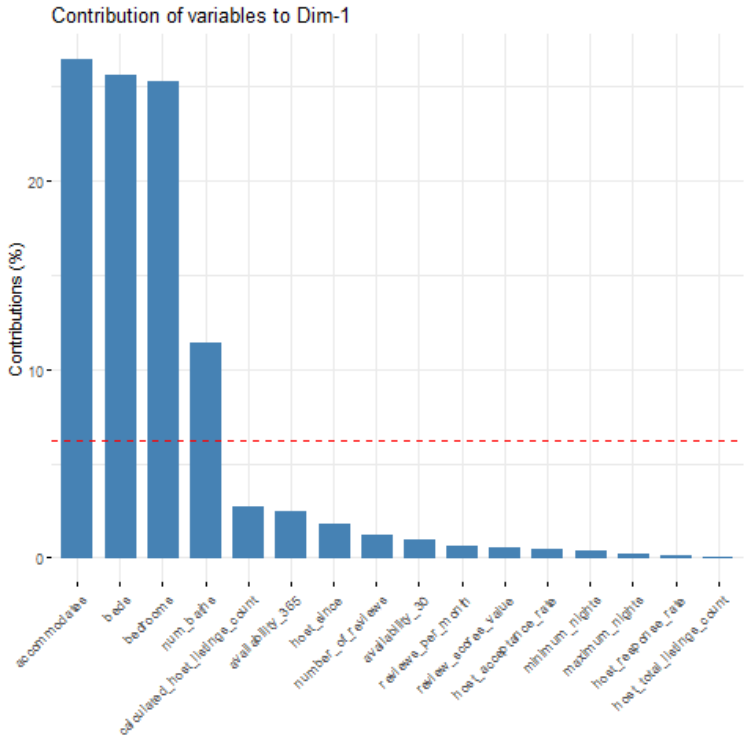


Image 25: Contribution of variables to PC1

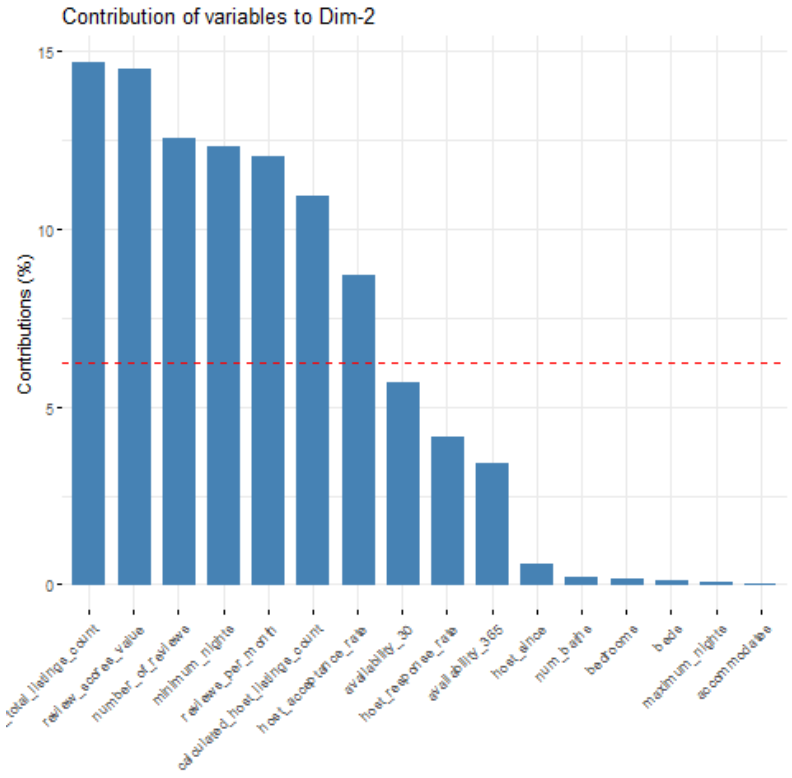


Image 26: Contribution of variables to PC2

Now we will observe the contributions of the variables to PC1 and PC3:

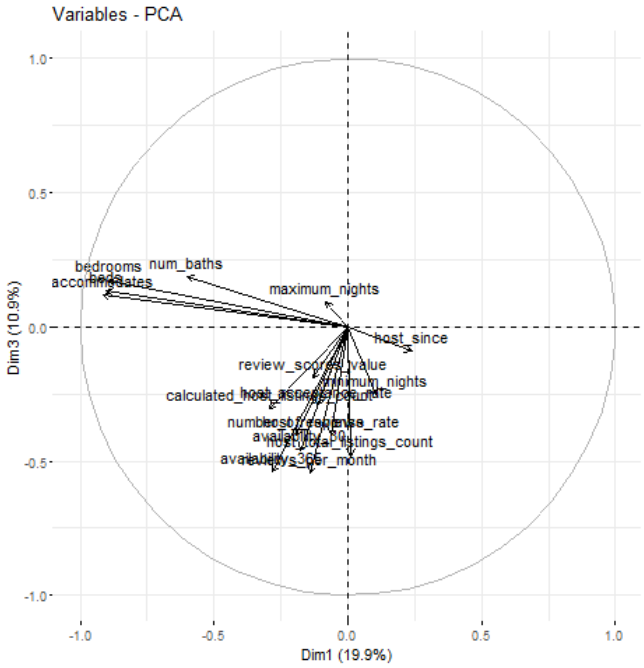


Image 27: PCA Variables factor map (PC1 and PC3)

It is a bit hard to distinguish the different names on the plot but as shown in the barplot below, the variables with a high contribution to PC3 are *reviews_per_month*, *availability_365*, *host_total_listings_count*, *availability_30*, *host_response_rate* and *number_of_reviews*. PC3 is explained by various variables.

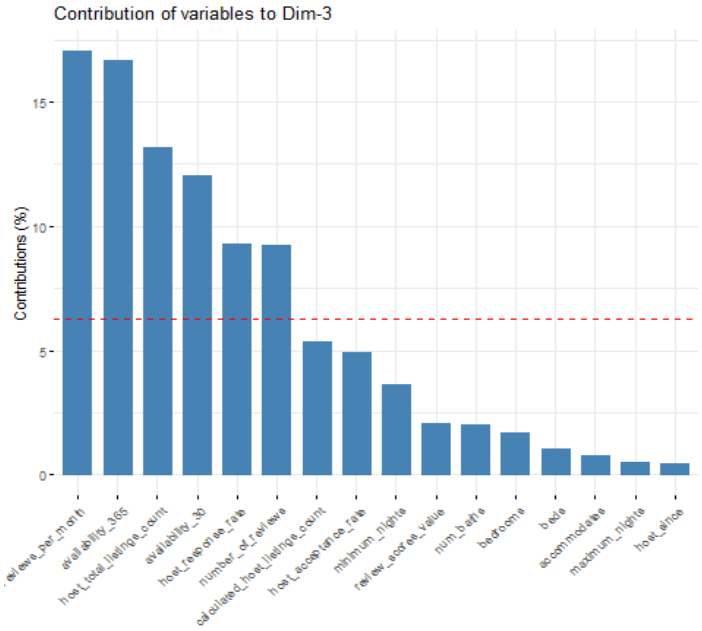


Image 28: Contribution of variables to PC3

Finally, the factor map which shows the contribution of the variables to PC1 and PC4 is the one below.

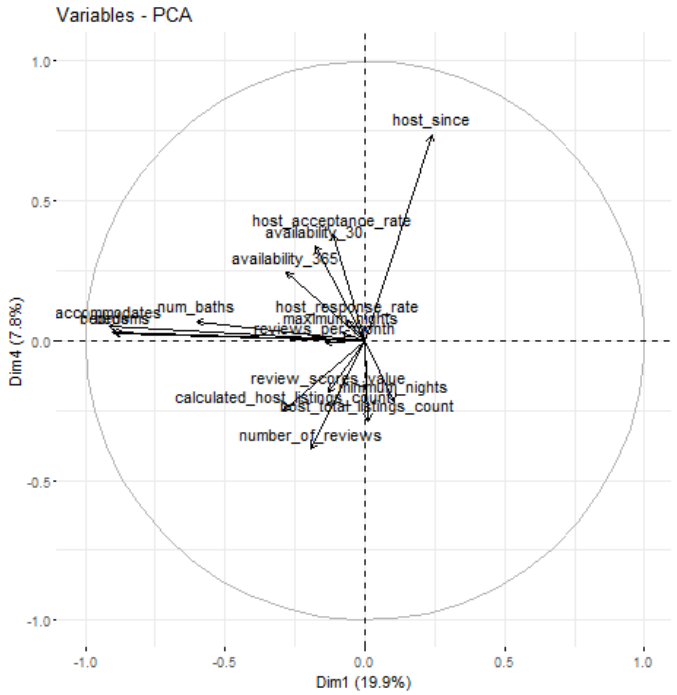


Image 29: PCA Variables factor map (PC1 and PC4)

In this case, it is easy to see that host_since explains most of PC4 (more than 40% to be exact). It can actually be seen that its corresponding arrow is very large and quite close to the PC4 axis. In addition, *number_of_reviews*, *host_acceptance_rate*, *availability_30* and *host_total_listings_count* also contribute to it but in small numbers as shown in the following plot.

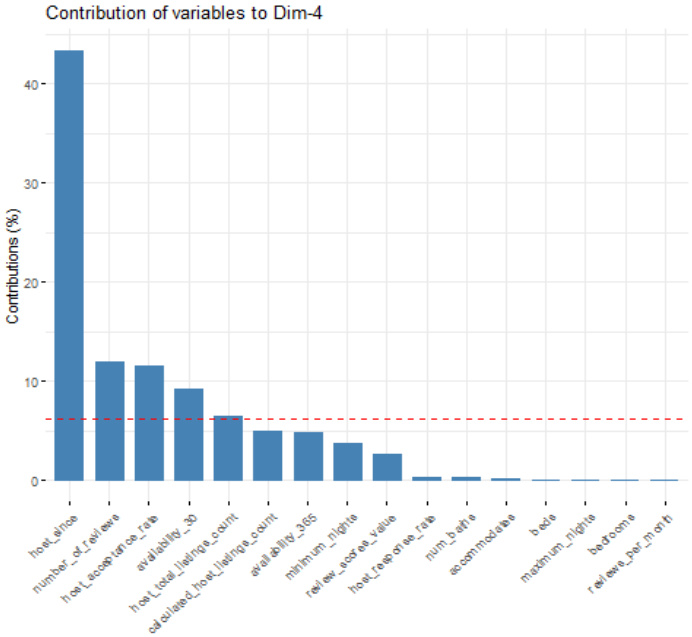


Image 30: Contribution of variables to PC4

We will not show here the contributions of the variables to PC5 and PC6 because their percentage of explained variance is lower than the already shown and we have not found any highlighting or meaningful results from them. However, if the reader is interested, they can be found in the Annex.

Now we will show the most interesting plots found throughout the PCA analysis. We will mainly expose 2D results obtained from PC1 and PC2, the most relevant principal components.

Before diving into the results, it is important to note that PC1 and PC2 together only explain up to 35,5% of the total variance. Therefore, what we will discuss from this moment on will not be conclusive but it will give us small hints on how the different variables are related.

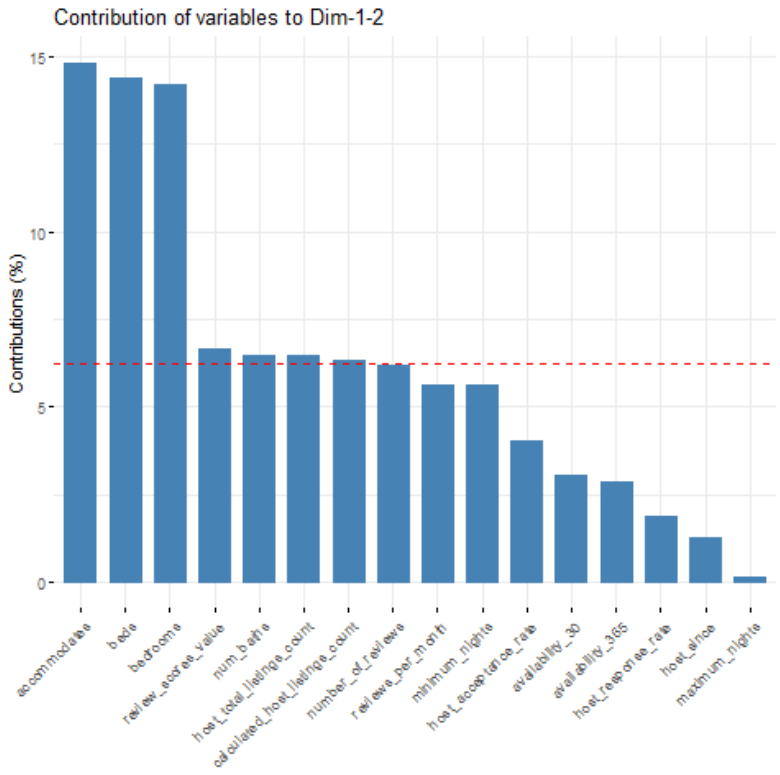


Image 31: Contribution of variables to PC1 and PC2 together

When interpreting the plots below, it will also be important to keep in mind that they are mainly explained by the following variables: *accommodates*, *beds*, *bedrooms*. But *review_scores_value*, *num_baths*, *host_total_listings_count* and *calculated_host_listings_ _count* also have a little but relevant influence on them.

The factor map of the individuals per se did not give us much information. Therefore, what we did was use supplementary variables in order to extract significant information. The factor map of individuals is the following:

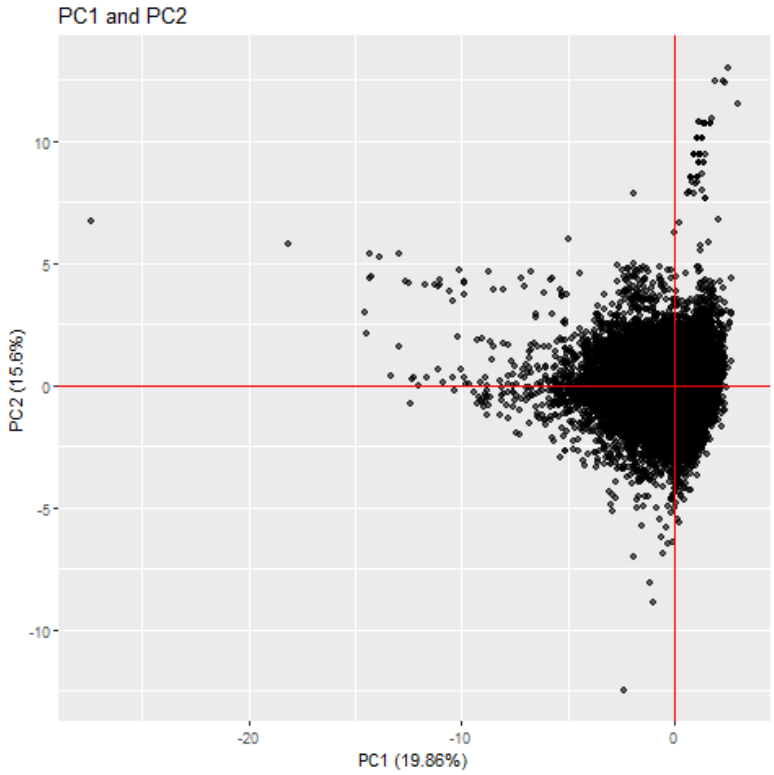


Image 32: Individuals factor map according to PC1 and PC2.

The huge black cloud of points in this plot is telling us that according to PC1 and PC2 most of the listings have a very similar behaviour and follow a similar pattern. However there are some small groups of listings which behave a bit differently.

In the first quadrant we can see a small group distributed through the PC2 axis and a bit far from the big cloud. According to Image 32 (variables factor plot PC1 and PC2) they are probably listings with few and bad reviews and their host has many other listings and with a high number of minimum nights.

In the second quadrant there is no such obvious group as in the first one but the listings that are a bit far from the main group probably admit a high number of guests.

Finally for the third quadrant, the ones away from the black cloud are likely listings with many reviews and whose hosts are responsible for a low number of listings.

In order to contrast the stated hypotheses and get more information from the PCA, we repeated this plot adding different supplementary variables by colouring the listings according to the factors of these additional variables. Let us now introduce the obtained results.

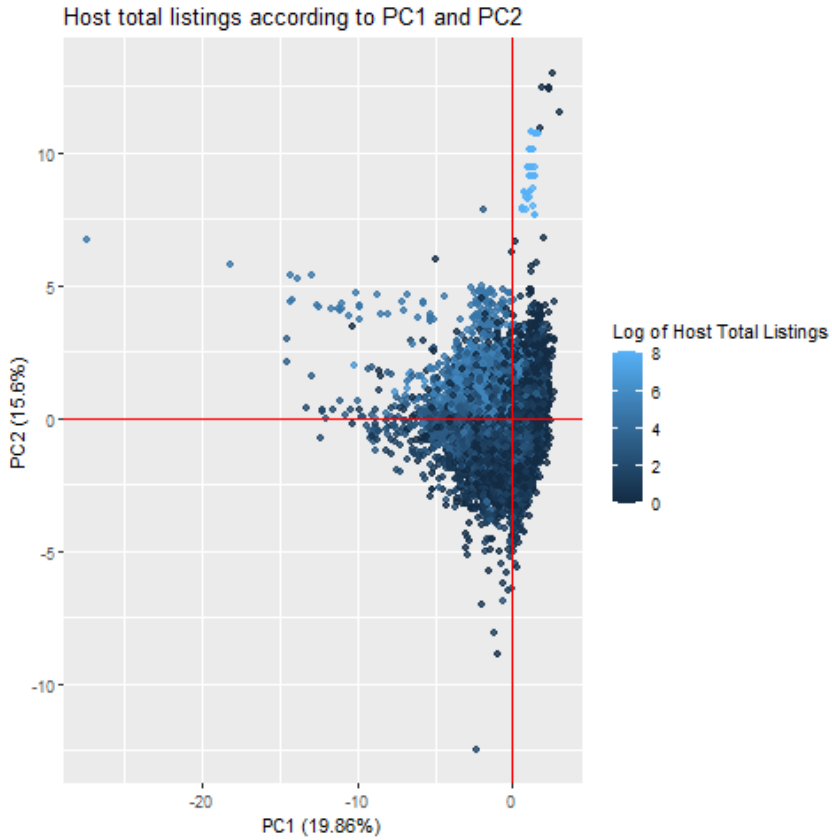


Image 33: Individuals factor map according to PC1 and PC2 with host_total_listings_count as supplementary variable.

Adding host_total_listings_count as a supplementary variable we can see that the group in the first quadrant, as mentioned in the previous page, are mainly listings whose host owns a large number of Airbnbs. This little cloud seems to be a group of individuals that have similar qualities, but we will be able to see if it's true in the later plots.

Nonetheless, we can also find some individuals in that area for which the host total listings count is low, just one.

In the second quadrant, we can find listings whose host owns quite a lot of Airbnbs. There seems to be a little group (in blue-grey colour) in the second quadrant with hosts with a high number of listings and then another one closer to the centre whose hosts own even more listings though not as much as the listings in the first quadrant.

As for the third quadrant, this variable is not able to explain the observed dispersion and does not provide any relevant insights on the listings in the fourth either.

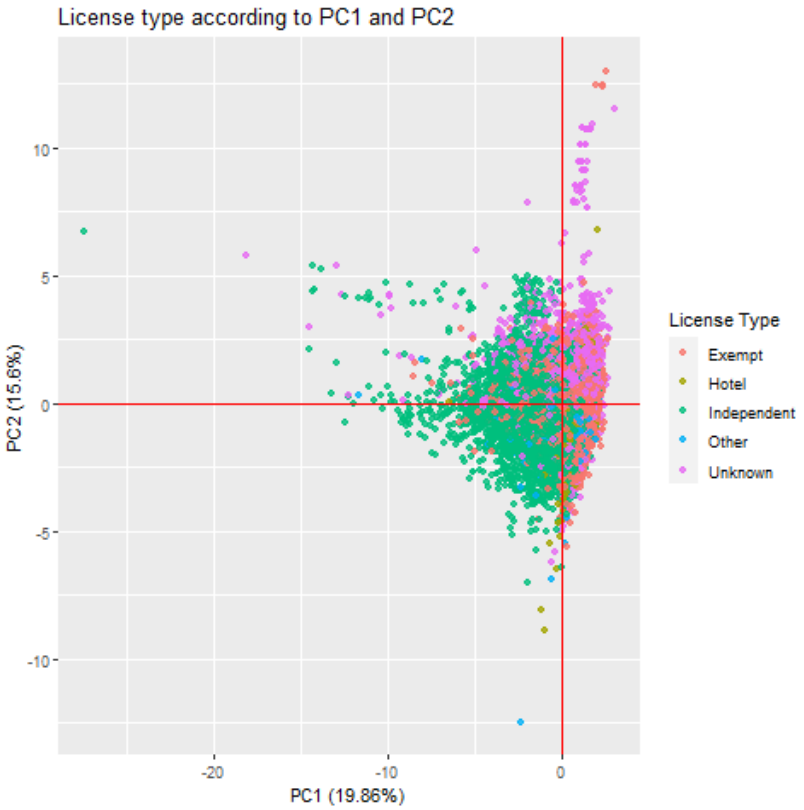


Image 34: Individuals factor map according to PC1 and PC2 with license as a supplementary variable.

Applying *license* as a supplementary variable provides us some relevant insights. By taking a look at (Image 34: variables factor map PC1 and PC2) we can see the following:

On the one hand, we can find most of the listings with an unknown license in the first quadrant. Therefore it seems that listings with a higher value for *minimum_nights* are the ones with an unknown type of *license*, which is not surprising because it is likely that they don't have one. Listings without a license must be rented for a minimum of 32 nights due to Catalunya tourist regulations.

Then the ones with an independent *license* can be mainly found in the second and third quadrants, telling us that listings which can accommodate more people are usually more available and those that have good hosts tend to have an independent *license* type.

Finally, listings exempt from licensing and others are mainly found in the fourth quadrant.

The following plot shows that listings with a high value of *min_nights*, particularly with *min_nights* >= 30 can be found in the first and second quadrant.

It is interesting to see that the little group of listings not contained in the main cloud of points that can be found in the first quadrant only admit long term rentals. Also, it seems listings with an unknown license only admit long term rentals.

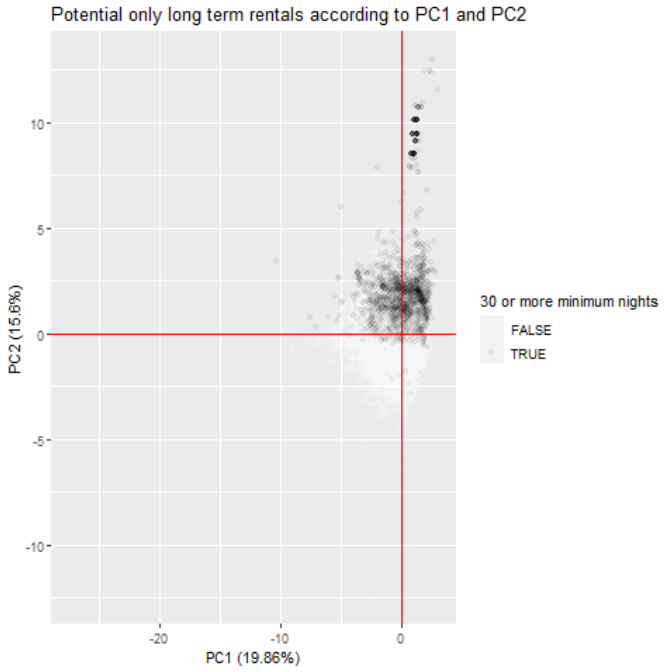


Image 35: Individuals factor map according to PC1 and PC2 with min_nights>30 as a supplementary variable

Let us now see what happens when we add *neighbourhood_group_cleansed* as supplementary.

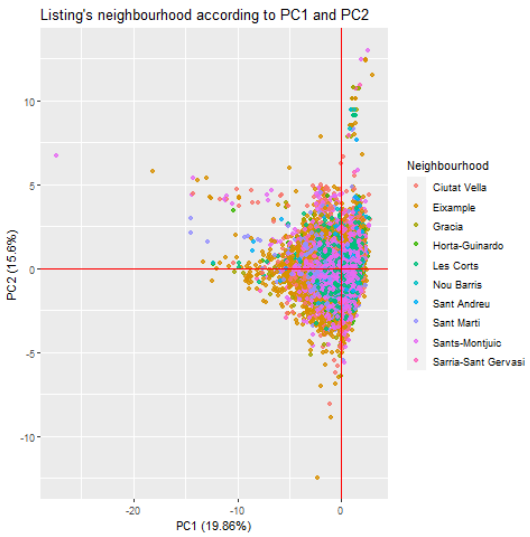


Image 36: Individuals factor map according to PC1 and PC2 with listings_neighbourhood as a supplementary variable

In this case it can be seen that most listings from Eixample and Sant Martí are in the second and third quadrants.

To have a better view of the stated listings we constructed the following plot:

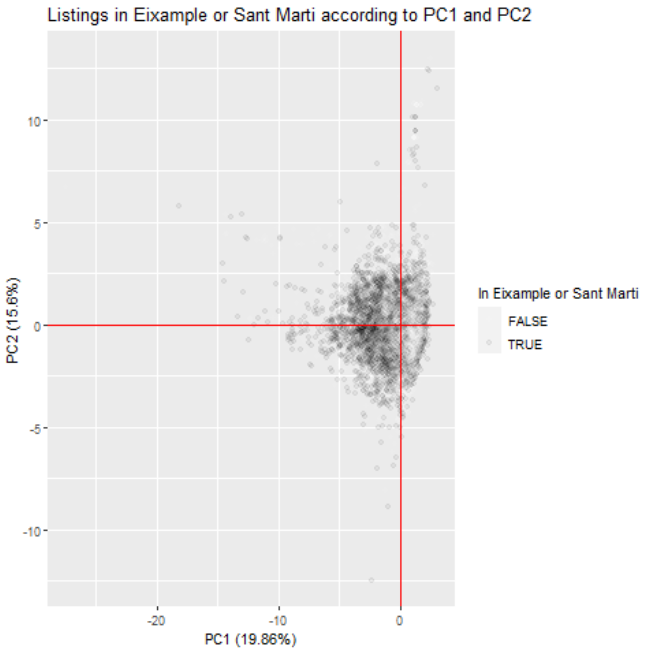


Image 37: Individuals factor map according to PC1 and PC2 with 'listing in eixample' as a supplementary variable.

By the way they are positioned in the plot, it seems they coincide with listings that accept a high amount of guests.

In order to prove this statement we did a biplot (Image 38) showing the average number of accommodates accepted in each neighbourhood. As expected, the Eixample and Sant Martí neighbourhoods are the ones with the highest average (4).

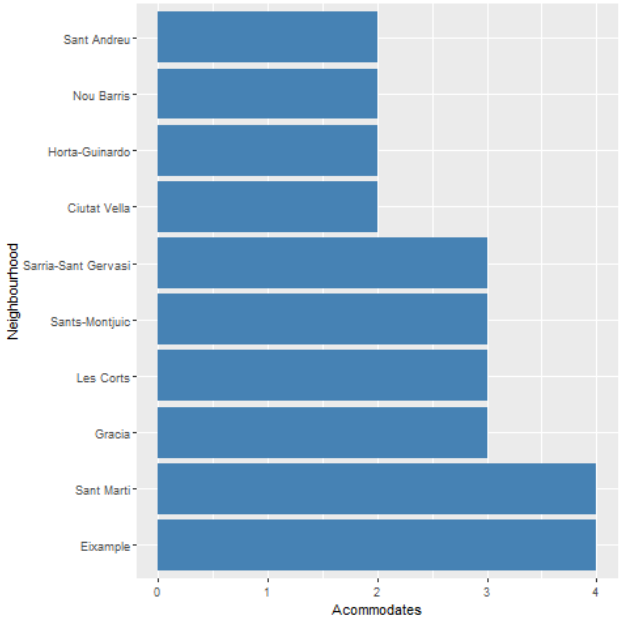


Image 38: Bivariate plot showing the average of accommodates per neighbourhood.

The next plot is coloured depending on the amount of people a listing accommodates. We know that the variable *accommodates* makes a considerable contribution to PC1, and this is what we can see here: the ones on the left side of the plot accept a low number of accommodates and the ones on the right side the opposite. This plot also shows that listings located in Eixample and Sant Martí are actually the ones that accommodate the most people.

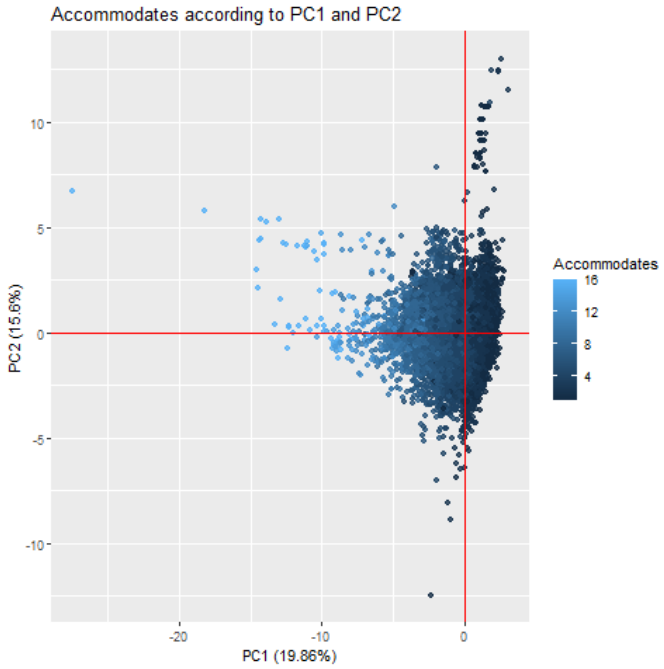


Image 39: Individuals factor map according to PC1 and PC2 with accommodates as a supplementary variable.

Now, take a look at the following graph:

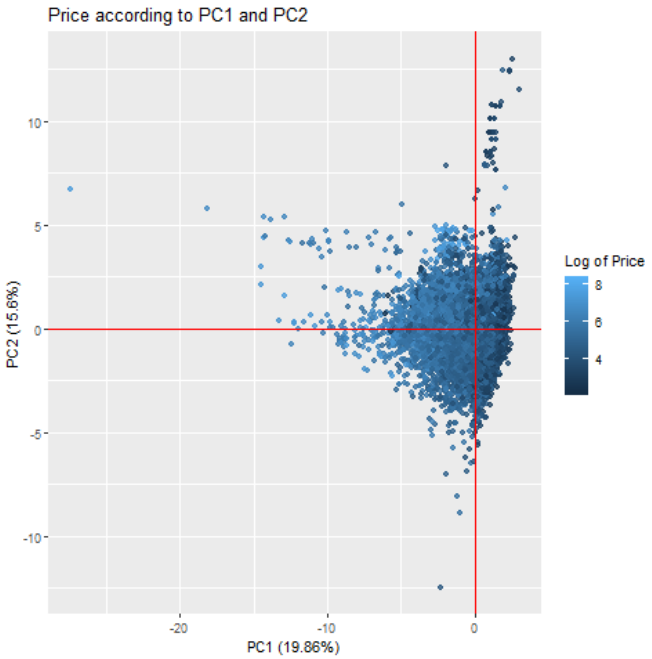


Image 40: Individuals factor map according to PC1 and PC2 with price as a supplementary variable.

There is a clear positive relationship between the price of the listings and the amount of accommodates they accept. There is a little group of listings which are quite expensive in the second quadrant close to the main cloud.

In addition, note that the listings with high prices coincide mainly with those in Eixample, to see it more clearly check the following image.

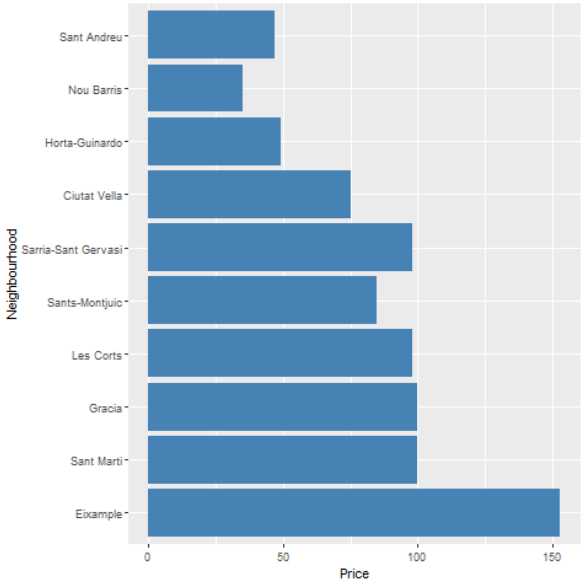


Image 41: Bivariate plot showing the average price of the listings per neighbourhood.

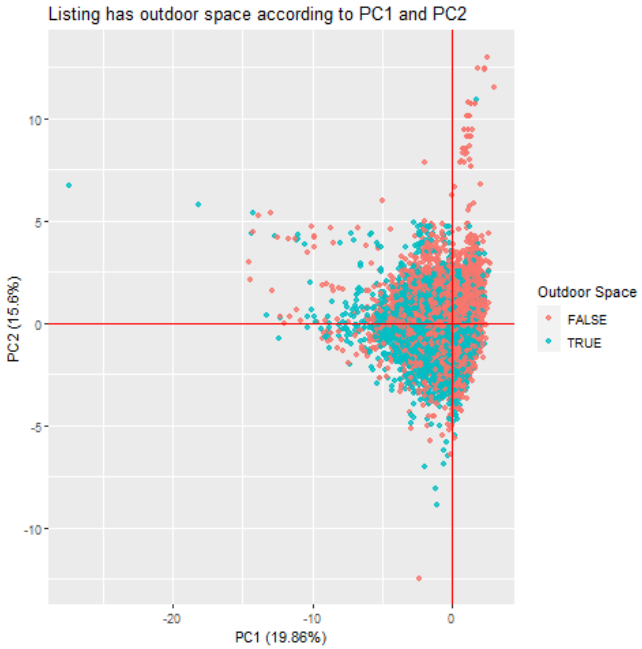


Image 42: Individuals factor map according to PC1 and PC2 with 'listing has outdoor space' as a supplementary variable.

The fact that a listing has or not an outdoor space is not deterministic. However, if we take a look at the Image 34 it seems that the ones that do not have an outdoor space mostly have an

unknown type of license whereas the ones with an outdoor space seem to coincide with the ones with an independent license type. Note that there are some exceptions.

Image 43: Individuals factor map according to PC1 and PC2 with 'host is superhost' as a supplementary variable.

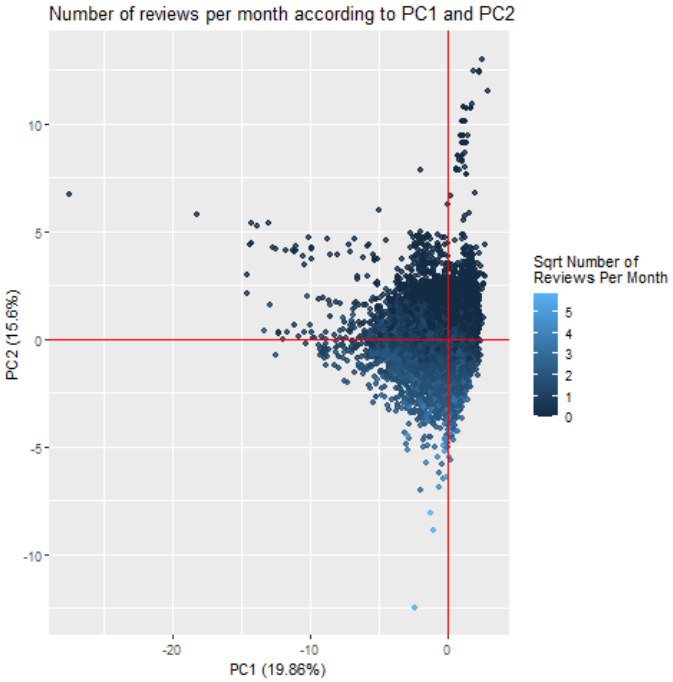


Image 44: Individuals factor map according to PC1 and PC2 with number_of_reviews as a supplementary variable.

From the two plots above we can see that in general superhosts tend to have more reviews. However, there exist listings with many reviews whose host is not a superhost. Also the outliers in the third quadrant seem to be listings with a very large number of reviews.

Let us now see what happens when we add *review_score_value* as a supplementary variable.

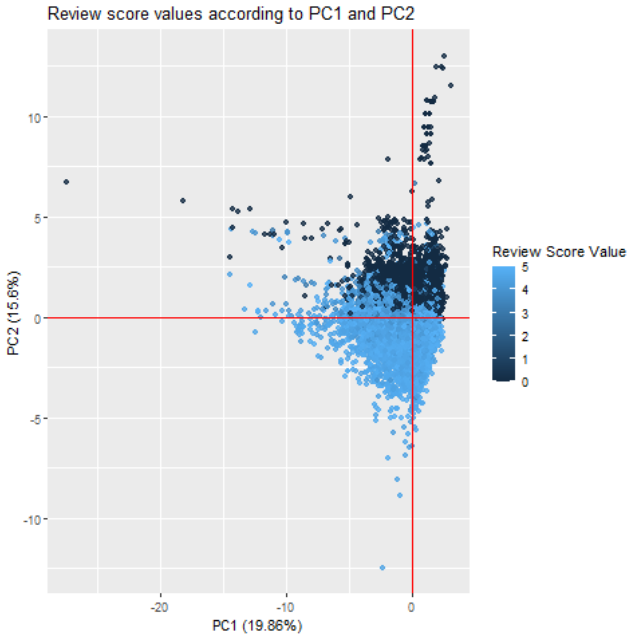


Image 45: Individuals factor map according to PC1 and PC2 with review_score_value as a supplementary variable.

First of all, note that the review_score_value shows if a guest agrees or disagrees with the price of the listing and guests usually tend to either absolutely agree with the price or absolutely disagree with it. Be aware that some of the listings with a zero score did not have any reviews and were imputed.

We can see that superhosts tend to have higher review scores and also it seems that those listings whose hosts own a lot of listings and have a minimum nights stay of 30 or more are the ones people tend to review badly (in terms of price) or not review.

Image 46: Individuals factor map according to PC1 and PC2 with property_type as a supplementary variable.

The previous plot clearly shows, according to Image 28 that the listings that are a private room in a rental unit are the ones that admit a low number of people. In other words, the property type variable is related to PC1. Indeed, the entire rental units and ‘others’, are the ones that accept more guests.

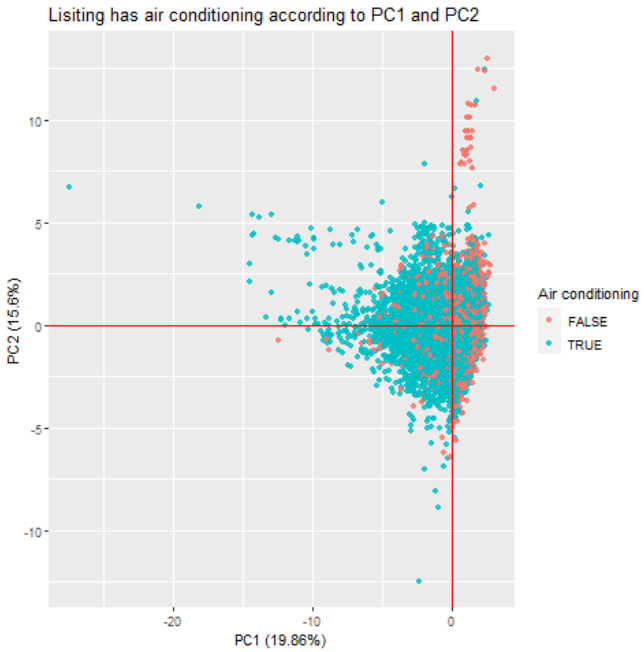


Image 47: Individuals factor map according to PC1 and PC2 with ‘air conditioning’ as a supplementary variable.

Surprisingly this plot is a bit similar to the previous one. It seems that most listings do have air conditioning but that the ones that do not are mainly the private rooms in rental units.

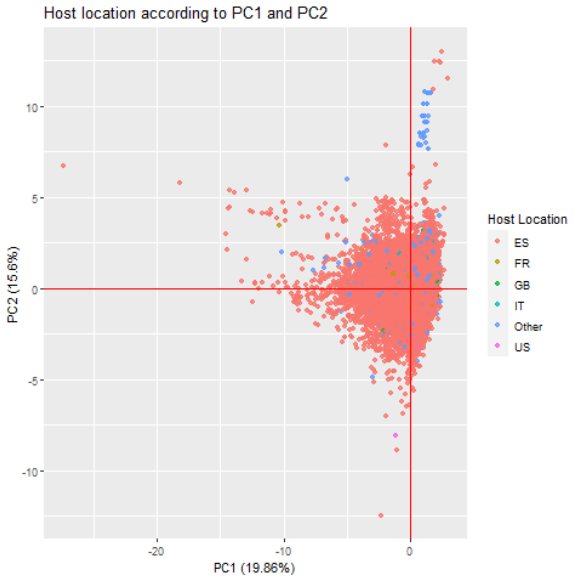


Image 48: Individuals factor map according to PC1 and PC2 with property_type as a supplementary variable.

With this plot it can be seen, as already known, that most of the hosts are from Spain. But the relevant information is that the little group of listings painted in blue have hosts, probably the same one, who are not from Spain.

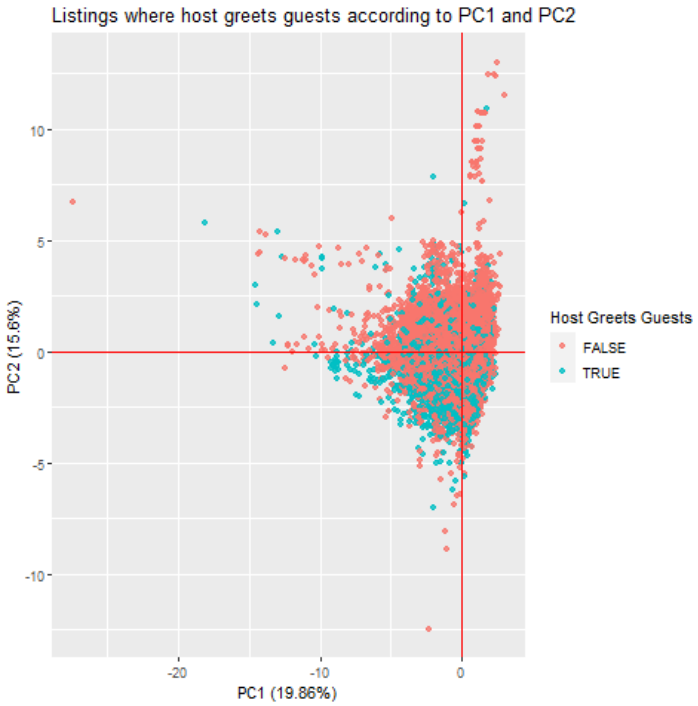


Image 49: Individuals factor map according to PC1 and PC2 with *host_greets_guest* as a supplementary variable.

The interesting insight from this last plot is that, in general, when the host greets the guests the listings tend to have a lot of reviews and higher review scores. However, there exist listings with high reviews whose host does not greet guests.

In addition, if you take a look at the Image 43 you will see that it looks very similar to this one, which tells us that usually superhosts are the ones that greet guests and vice versa, though, as always, there are some exceptions.

10. MCA analysis

As happened in PCA, first of all we had to keep some of the categorical variables in order to reduce the amount of final dimensions and also to have a better performance from de MCA.
From the 29 variables we can find in our dataset, only 13 were used as active variables for the MCA.
We kept the 13 variables that at first sight could be more powerful in order to detect groups and see some patterns considering our target variable (price). This variables are the following: *neighbourhood_group_cleansed, property_type, wifi, longTermStays, hairDryer, aircon, heating, tv, hostGreets, outdoorSpace, parkingOnPremise, pool, bbq*

The MCA analysis was done using the Logical Table method. Taking this into account, in order to keep the most important dimensions, we only had to consider those with an eigenvalue greater than 1/p with p as the number of active variables. In our case this value was 1/13=0.0769. Thus, our MCA analysis was done with 9 variables as the 10th had a value less than 0.0769. Moreover, these 9 dimensions explain more than the 50% of the variance of the data. In this part of the report only the plots from dimensions 1-2 / 1-3 will be inserted. You may find more plots in the Annex.

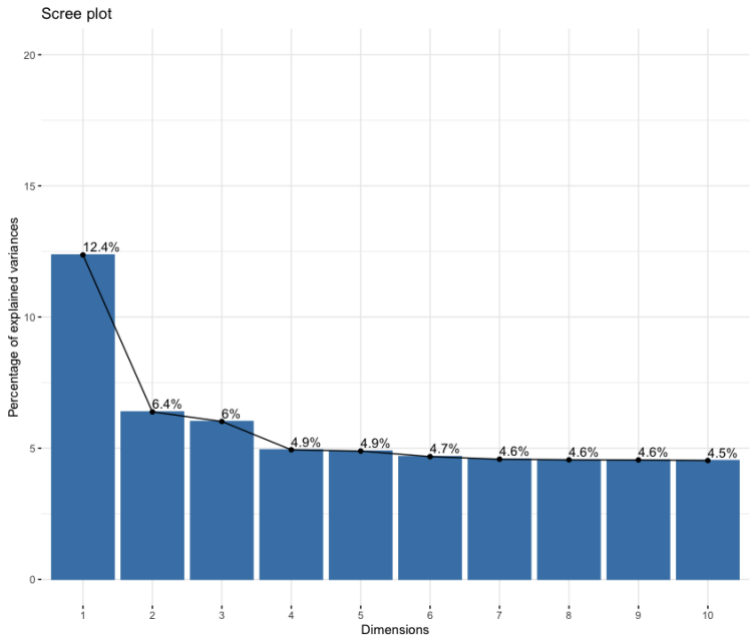


Image 50: Variance explained by dimensions 1 to 10

The plot above shows the percentage of explained variances of every dimension. We can find a stability elbow in the 4th dimension approximately. Despite this, the first 9 dimensions are used for the MCA.

As a first analysis of the MCA, we checked the individuals and features in 2 dimensions.

Image 51: Biplot variables / individuals in dimensions 1 and 2

The above biplot compares Dim1 vs Dim2. We can see that those individuals located more on top of the plot are listings located in Sarrià-Sant Gervasi or Sant Martí. Listings with luxury amenities such barbecue, pool and private parking. On the other hand, those located on the right are more simple listings, without wifi nor tv nor heating and are rooms in rental units.

Image 52: Biplot variables / individuals in dimensions 1 and 3

Comparing now Dim1 with Dim3 we can see that in Dim3 those listings located at the bottom of the plot are listings where the host greets you. On the other hand, those at the top are other types of properties and may be situated in Sarrià-Sant Gervasi, Les Corts or Sant Martí.

After this first analysis, we wanted to know what explained each dimension. In the following image we can see variable's contribution in each dimension:

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8	Dim 9
neighbourhood_group_cleansed	0.069477425	0.2385017275	0.068668104	4.736843e-01	0.3614001520	8.375932e-01	9.416842e-01	0.9901908988	9.943756e-01
property_type	0.523308920	0.1755556303	0.293138562	1.034888e-02	0.1058149449	2.182695e-02	4.948404e-03	0.0021355281	8.336735e-04
wifi	0.058754519	0.0002089333	0.100982702	2.155281e-02	0.2657647256	1.275945e-03	2.857829e-03	0.0027543892	1.076067e-03
longTermStays	0.077980151	0.0234182295	0.025197743	6.333062e-02	0.1464344220	1.264431e-01	1.850689e-02	0.0018494595	2.337605e-03
hairDryer	0.332781588	0.0125511624	0.069150561	3.831207e-03	0.0394456184	2.026321e-04	1.163653e-02	0.0021058489	6.272169e-05
aircon	0.553494370	0.0289188667	0.040345217	4.980219e-04	0.0007374801	4.584514e-04	6.893464e-04	0.0002494127	1.459453e-04
heating	0.369381937	0.0019986774	0.017841167	1.233608e-05	0.0486652138	2.923977e-03	3.771928e-03	0.0007476199	1.596041e-03
tv	0.453078003	0.0253989411	0.022141465	1.398037e-03	0.0050542327	2.540647e-02	3.896009e-03	0.0004841490	1.961185e-04
hostGreets	0.027014211	0.1332819819	0.336887006	5.171665e-03	0.0325029482	8.426971e-03	5.212144e-05	0.0004495124	7.105235e-04
outdoorSpace	0.177666620	0.0436872979	0.084587178	8.640508e-02	0.0118113368	3.466227e-04	6.074809e-03	0.0007222206	1.894358e-04
parkingOnPremise	0.046272373	0.3816752424	0.012082799	3.864734e-02	0.0408514193	8.138845e-05	6.534939e-03	0.0005007334	2.057092e-05
pool	0.030693817	0.1859282230	0.248333399	3.851027e-02	0.0114970438	3.429123e-03	5.704300e-03	0.0001095856	6.969645e-05
bbq	0.001353505	0.1535083389	0.005082716	3.432165e-01	0.0061256590	1.028982e-03	1.551498e-03	0.0001758680	9.135577e-07

Image 53: Contribution of variables in each dimension

Taking the top 3 variables of each dimension, we can know what they represent:
Dimension 1: Explains air conditioning, type of property and tv
Dimension 2: Explains parking on premise, neighbourhood and pool
Dimension 3: Explains host greeting, type of property and pool
Dimension 4: Explains neighbourhood, barbeque and outdoor space
Dimension 5: Explains neighbourhood, wifi and long term stays
Dimension 6: Explains neighbourhood, long term stays and tv
Dimension 7: Explains neighbourhood, long term stays and hair dryer
Dimension 8: Explains neighbourhood, wifi and property type
Dimension 9: Explains neighbourhood, long term stays and heating

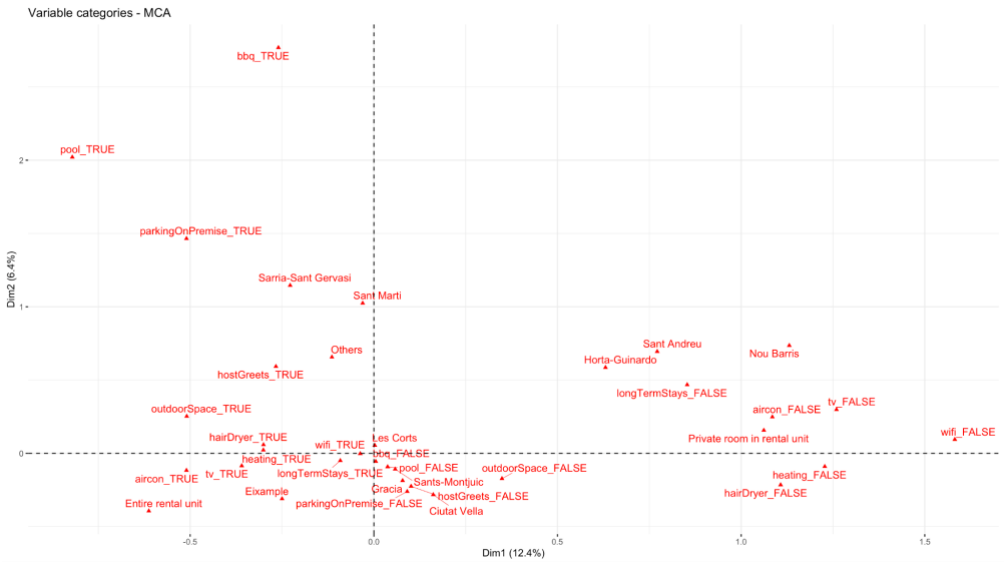


Image 54: Correlation between variables in dimensions 1 and 2

Let's take a look at the correlation between variables in dimensions Dim1 and Dim2. As we've seen, dimension 1 explains type of property and amenities while dimension 2 explains location and luxury amenities.

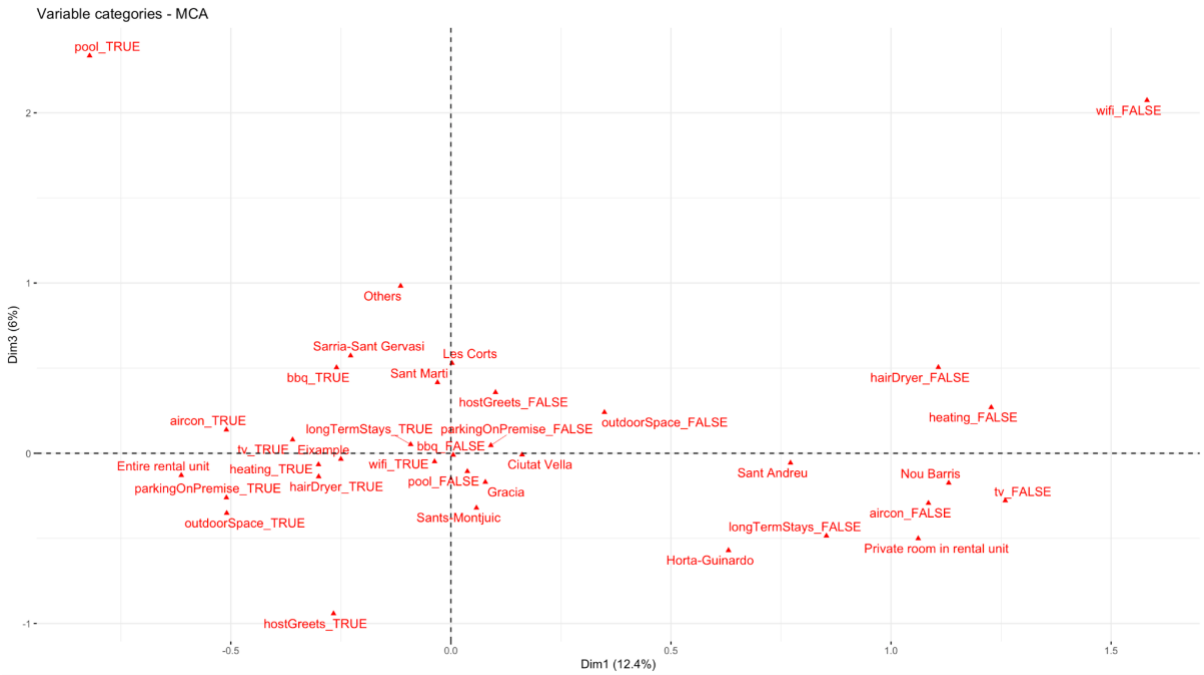


Image 55: Correlation between variables in dimensions 1 and 3

In the above plot we can see that dimension 3 explains almost the same as dimension 2: luxury amenities and location of the listings.

To make sure that these dimensions explain what has been discussed, we compute another MCA adding some categorical variables as supplementaries that can reinforce our explanation. These supplementary variables were *license* and *type_bath*:

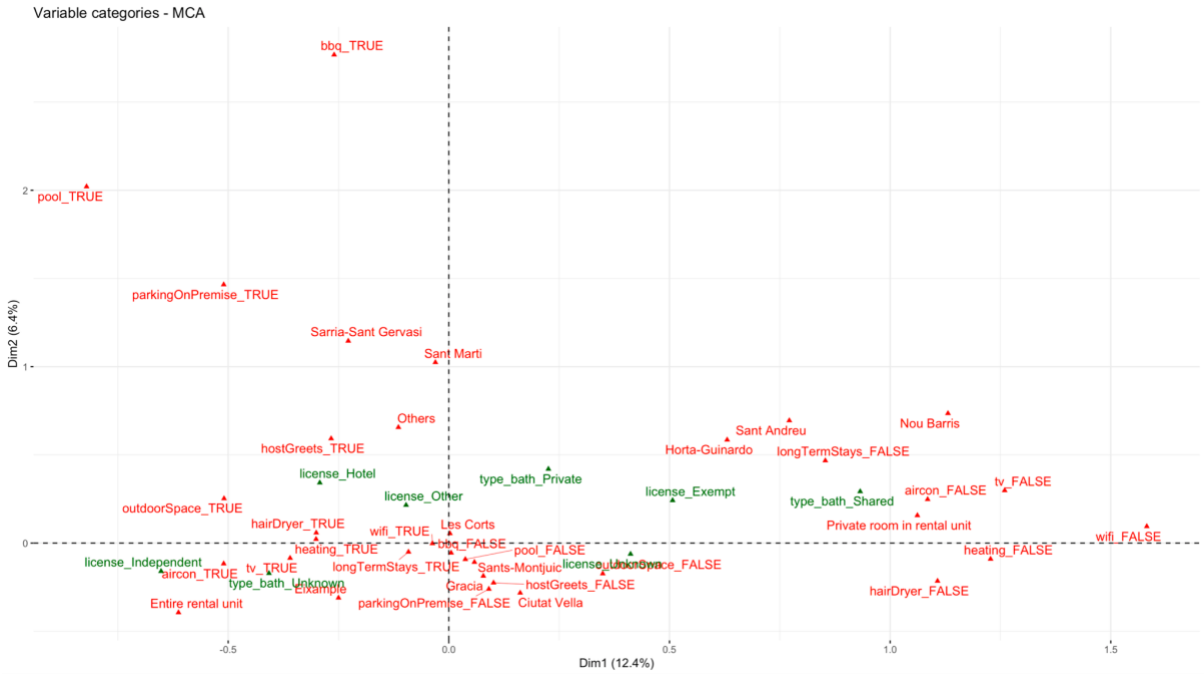


Image 56: Correlation between variables in dimensions 1 and 2 with supplementary variables

As we expected, those supplementary variables (green) added helped at the comprehension of the plot. In Dim1, those “bad” listings without essential amenities also have a bath shared. On the opposite site, those listings with good amenities have a hotel or independent license and its bath isn’t shared.

Additionally in our analysis, we can compute the contribution of the different modalities in each dimension.

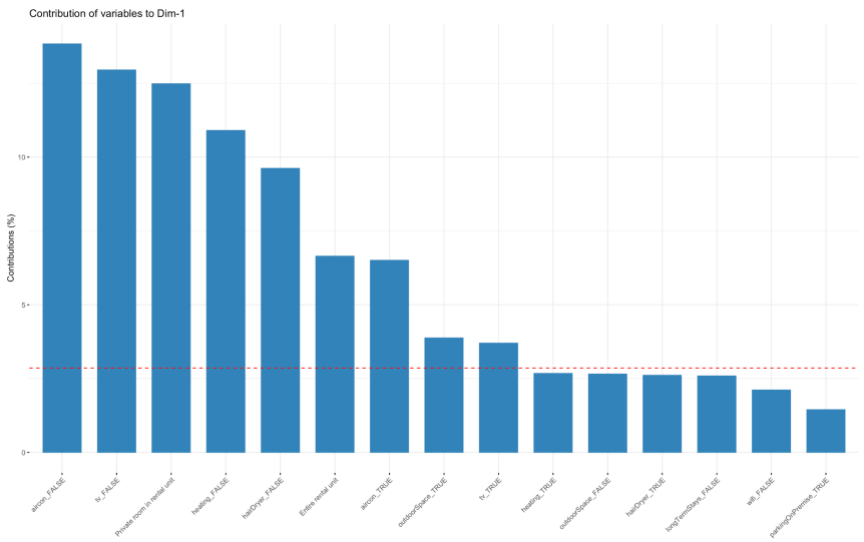


Image 57: Correlation between modalities in dimension 1

The most important modalities for Dim1 are those related to *aircon*, *tv* and *property type*

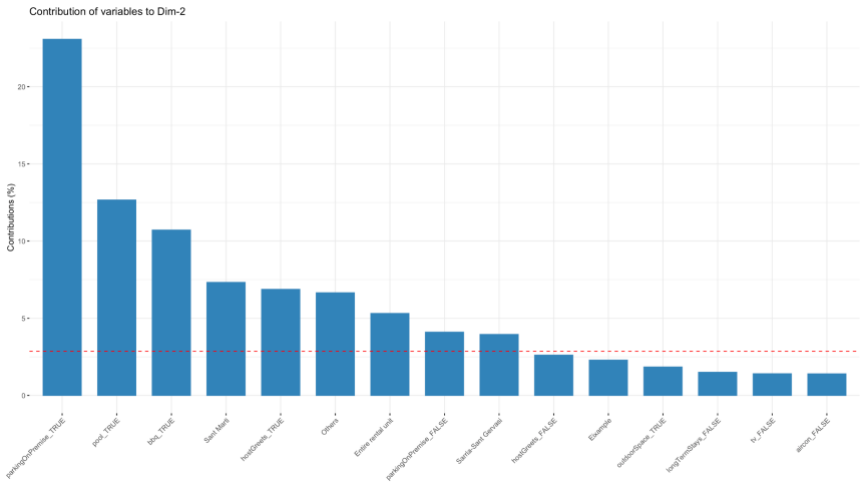


Image 58: Correlation between modalities in dimension 2

In Dim2, *parkingOnPremise* has a big contribution followed by *pool* and *Sant Martí (neighbourgood_group_cleansed)*.

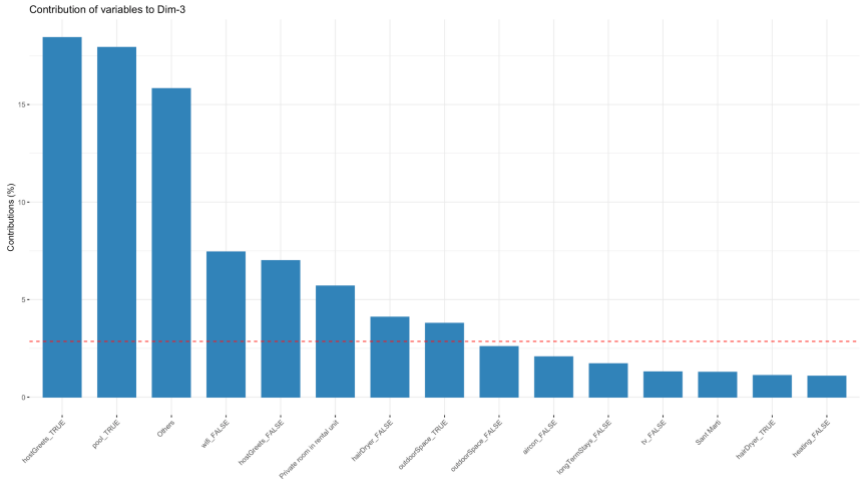


Image 59: Correlation between modalities in dimension 3

Finally, *hostGreets*, *pool* and *Others (property_type)* are the ones that contribute the most to Dim3.

After analysing the contribution and importance of the different variables and modalities to the MCA, we looked over the individuals. And we did it plotting them by variables, so we could look for clusters all over the variables. All plots will be inserted in the Annex.

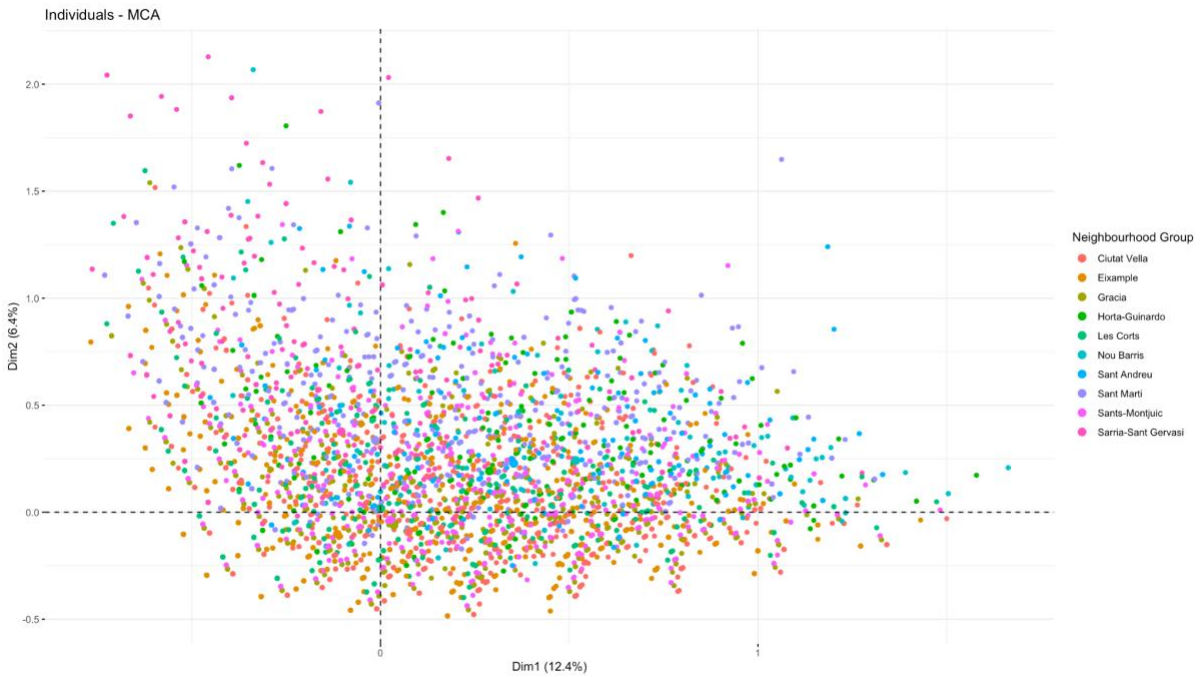


Image 60: Plot of individuals grouped by neighbourhood in dimensions 1 and 2

One of the most important variables is the neighbourhood group. At first sight with this plot it is impossible to detect any cluster or distribution as we have 10 modalities mixed all over the plot.

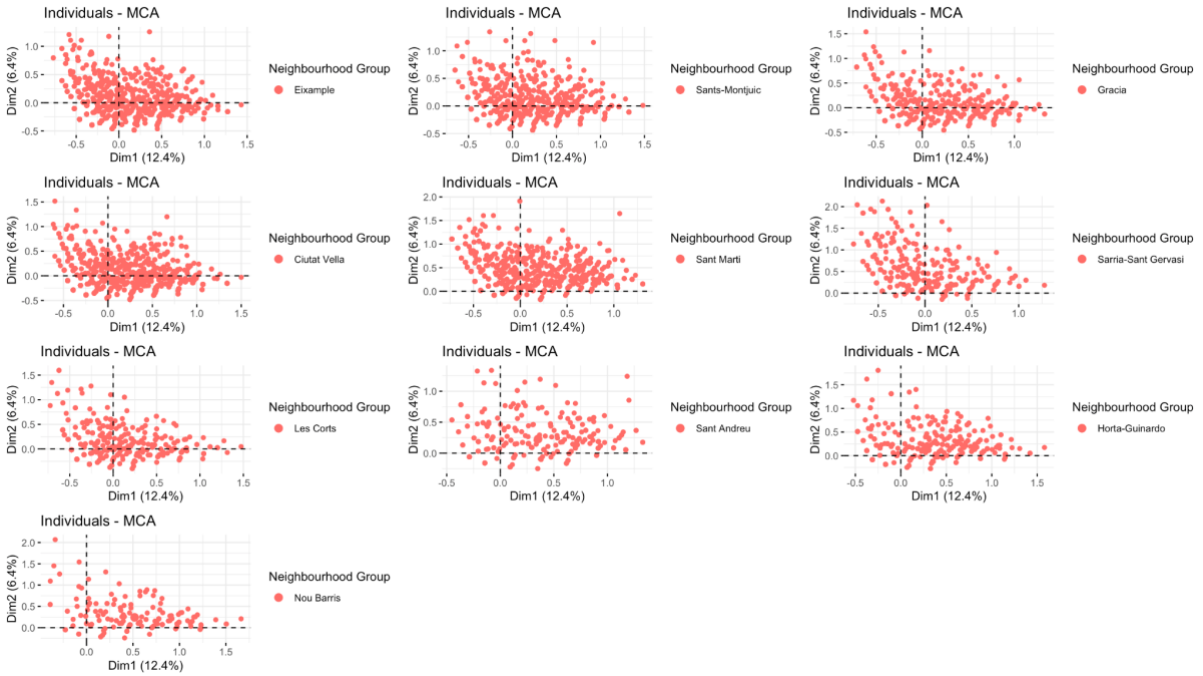


Image 61: Plots of individuals by neighbourhoods in dimensions 1 and 2

If we split modalities in 10 graphs we can see a really interesting plot in Nou Barris, where almost all the listings are located on the right side of the graph, what, as we saw before, means "bad" listings with no amenities, with a shared bath and with only a room rented.

Talking about the host's response time, as we will see, most of the hosts respond within an hour, independent of its location or amenities.

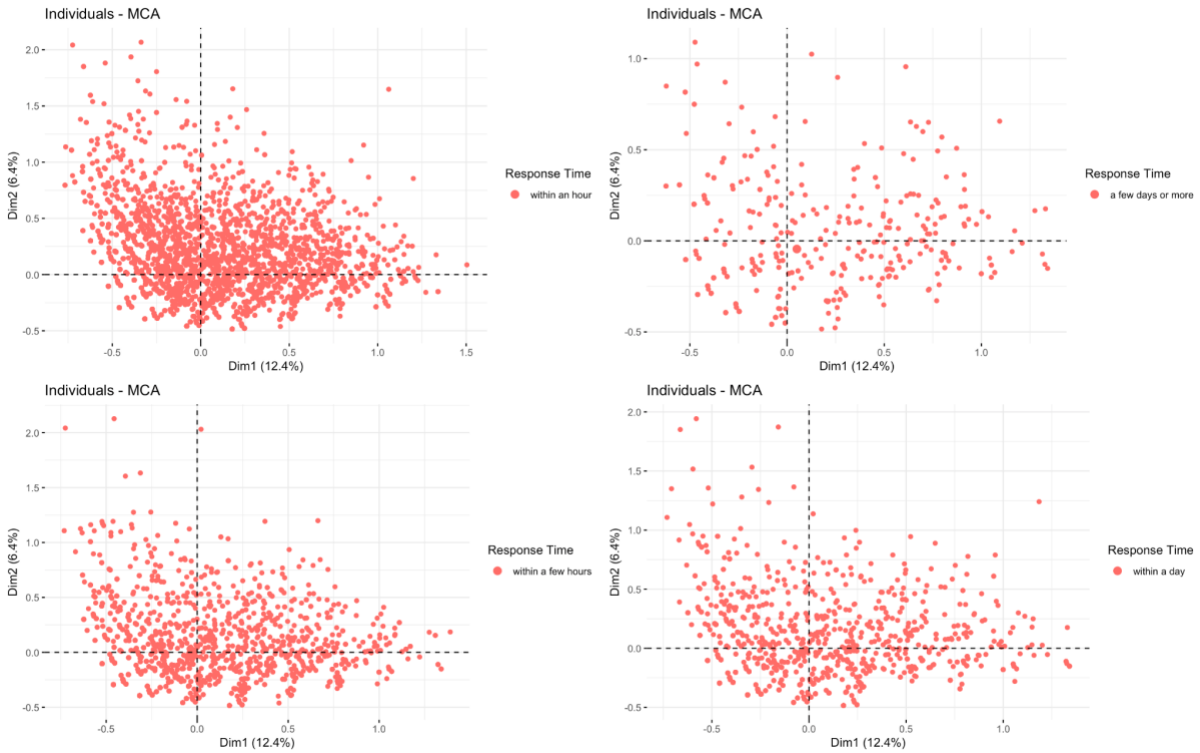


Image 62: Plots of individuals by response time in dimensions 1 and 2

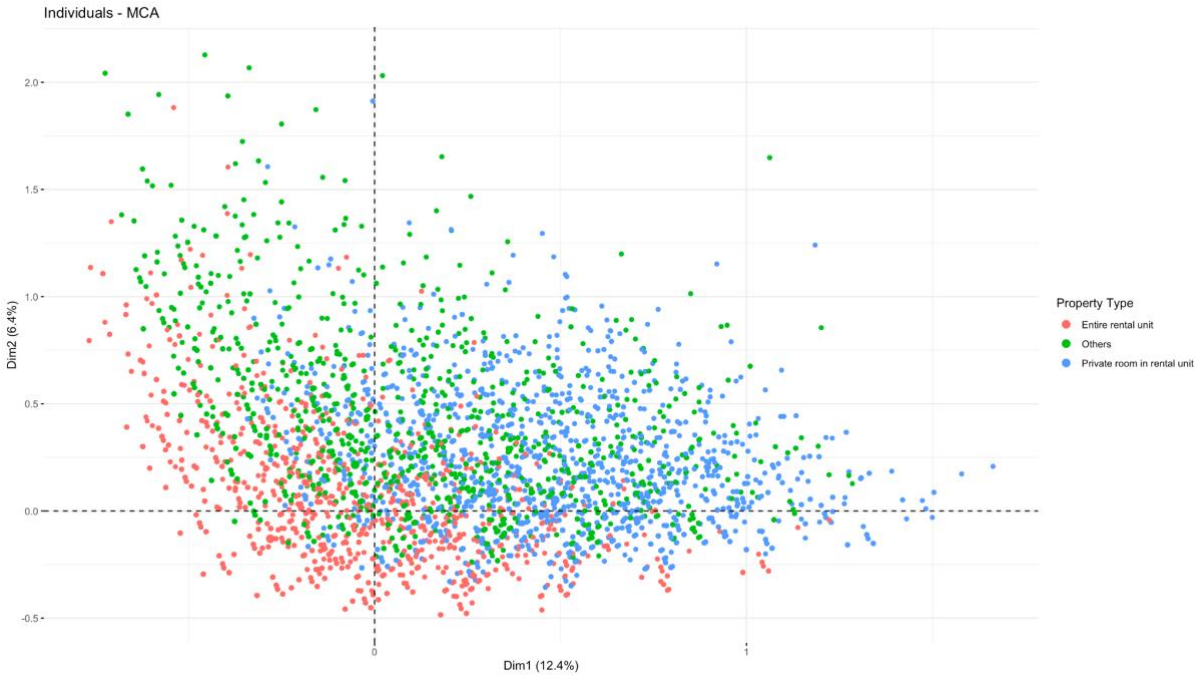


Image 63: Plots of individuals by property type in dimensions 1 and 2

In the plot above we can now distinguish between 3 groups. Listings with an entire rental unit are located mostly to the left side of the graph, private rooms in rental units on the right side and, finally, other types of properties between these two. In the first version of the dataset, these “Others” categories represented properties like boats, chalets, entire lofts… Properties

whose characteristics represent having good (or luxury) amenities and located in the best neighbourhoods.

Image 64: Plots of individuals by licence in dimensions 1 and 2

When it comes to licences, the splitted plot shows us that most of the listings from "Independent" licences are properties with good amenities and well located.

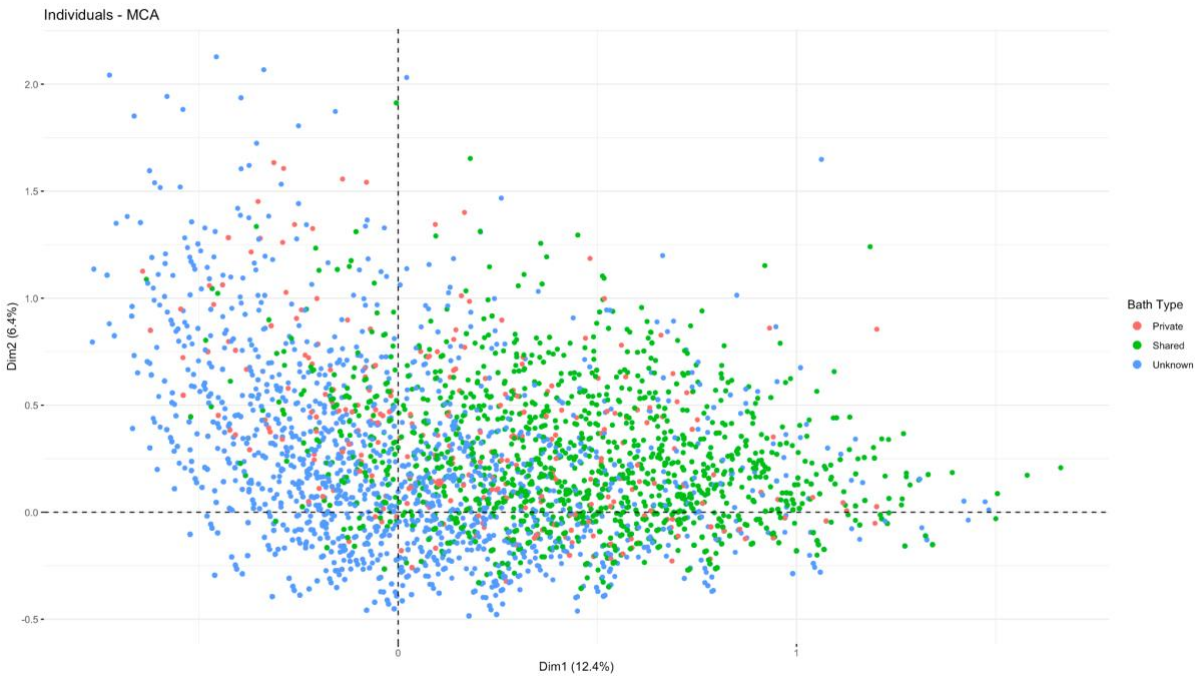


Image 65: Plots of individuals by bath type in dimensions 1 and 2

Type of bath shows a relationship between property types. Private rooms in rental units have a shared room. Blue dots are represented as “Unknown”. This is because the dataset didn’t specify if they were private or shared. Now, looking at the plot, we can assume that most of them were private baths.

Last but not least, amenities have shown really interesting plots.

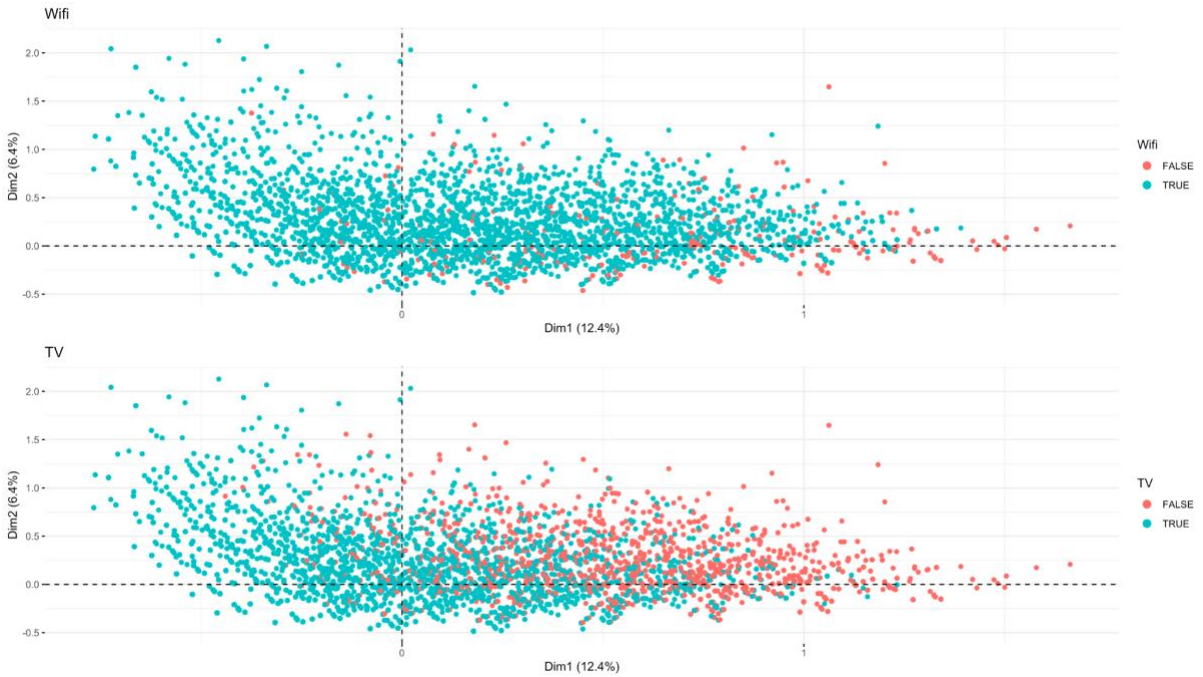


Image 66: Plots of individuals by WiFi and TV in dimensions 1 and 2

When dealing with TV and WIFI, we can see that those listings without any of both were the simplest listings, with rooms in rental units and located in neighbourhoods outside the city centre.

Image 67: Plots of individuals by outdoor space and parking in dimensions 1 and 2

Most of these listings don't have either an outdoor space nor a parking area as shown in the above plot. Neighbourhoods such as Sants, Gràcia or Ciutat Vella neither have parking as they are located more in the centre of Barcelona.

Finally, analysing the most luxury amenities (barbecue and pool), we can see that few listings can provide them:

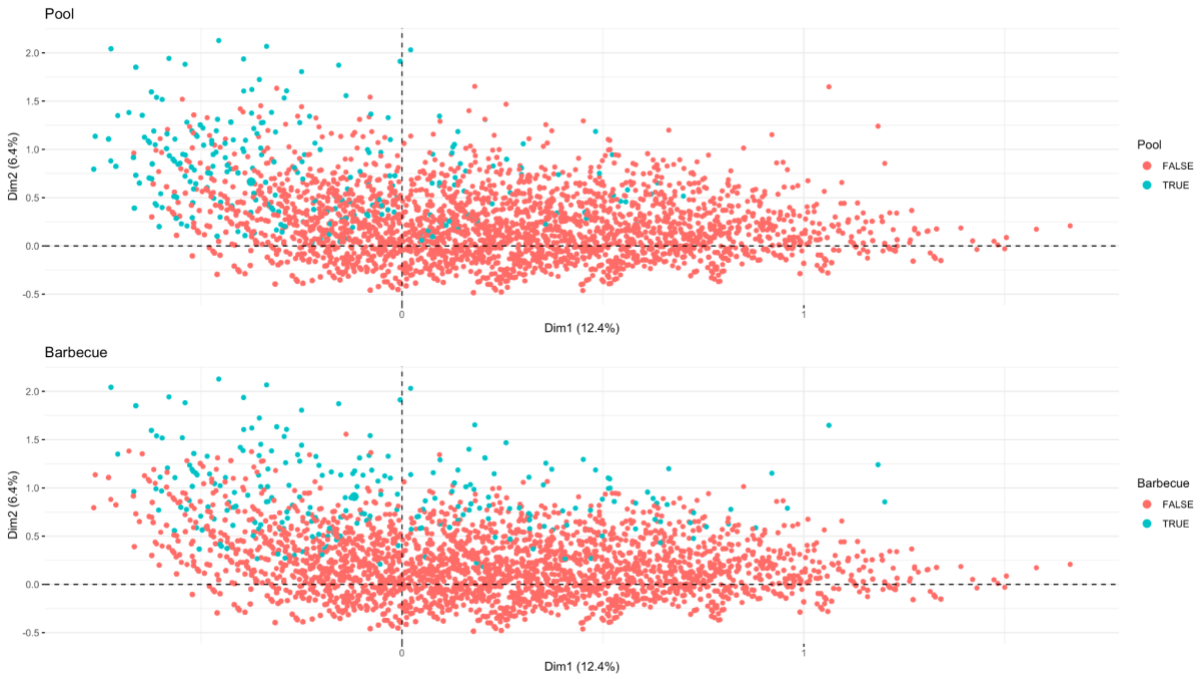


Image 68: Plots of individuals by pool and barbecue in dimensions 1 and 2

Actually, blue dot listings are located in neighbourhoods such as Sarrià-Sant Gervasi, Sant Martí, the most rich districts in Barcelona.

11. MFA analysis

The MFA analysis is using a combination of qualitative and quantitative variables in order to combine insights that PCA and MCA could give. In our PCA we have used 16 variables and in our MCA we have used 13 variables. In our MFA we have used 26 of the full 58 variables. Variables have been removed based on which ones we used in the PCA and the MCA, based on correlation among them and their usefulness for our MFA (We removed variables that did not have very useful extra information (e.g. if we keep the neighbourhood, we removed the latitude and longitude)).

The next step in MFA is to form variables into groups that belong together. However, in MFA a group can only be either categorical or numerical. Thus, some groups exists twice, once for categorical and once for numerical. We did that for the whole dataset and ended up with 8 groups.

The groups are the following:

Group name	Type	Variables
Host Information	*categorical*	host_response_time, host_is_superhost, host_identity_verified
Host Information	*numerical*	host_response_rate, host_acceptance_rate, host_total_listings_count
Location	*categorical*	nieghbourhood_group_cleansed
Physical Form	*categorical*	property_type, license, type_bath, hairdryer, aircon, heating, tv, hostgreets, parkingonpremise, pool, bbq
Physical Form	*numerical*	accommodates, num_baths, bedrooms
Booking Process	*categorical*	instant_bookable
Booking Process	*numerical*	availability_30, availability_365, minimum_nights
Price	*numerical*	price

Price is a supplementary variable as it is our target variable. We then run our MFA on the 26 variables and 8 groups with the price group as a supplementary variable and can observe the resulting Screeplot.

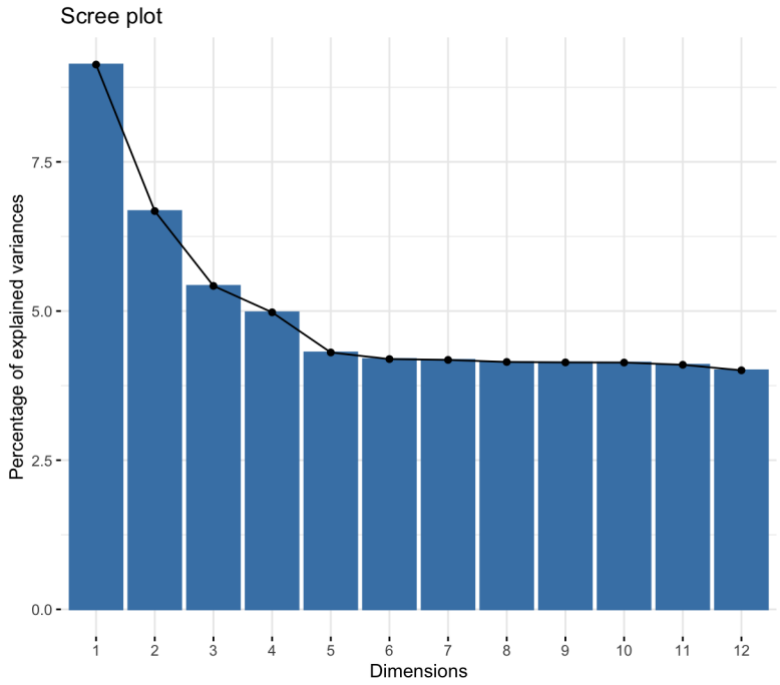


Image 69: Screeplot

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.211485	9.131502	9.131502
Dim.2	1.616753	6.675779	15.807281
Dim.3	1.313022	5.421635	21.228916
Dim.4	1.205738	4.978647	26.207563
Dim.5	1.042793	4.305825	30.513388
Dim.6	1.016017	4.195264	34.708652

Image 70: Eigenvalues of the dimensions and variance explained

Considering that we have 4 categorical and 3 numeric groups (plus one supplementary), we can see that our first dimension explains 9.1%, the second dimension 6.7%, the third dimension explains 5.42%, and the fourth explains around 5%. We can find a stability elbow after the 4th dimension approximately, with all the following dimensions contributing roughly the same explained variance. Therefore we decided to focus on the first four dimensions for the analysis, while considering that they will give us an interesting insight into our data, but will not be explaining our complete variance in the model.

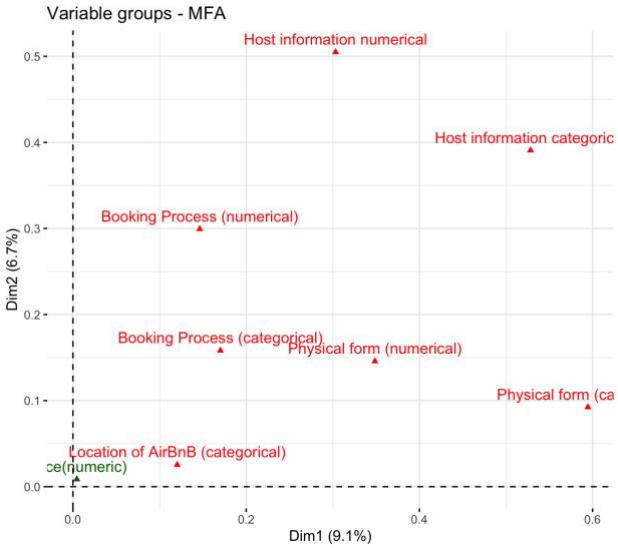


Image 71: Variable groups in MFA

First, we are considering the different variable groups and how they are contributing to our Dim1 and Dim2. As we can see Dim1 is mainly representing the physical form of the AirBnB. Moreover, the categorical host information is influencing it. This group includes the response time, superhost status and verification status of the host. Dim2 is mainly influenced by the numerical and categorical host information and the numerical booking process. This includes the availability of the AirBnB, the number of minimum nights required and the host acceptance rates, so could be summarised as the ease of booking. Interestingly enough, the location of the AirBnB is not really significant in our two first dimensions.

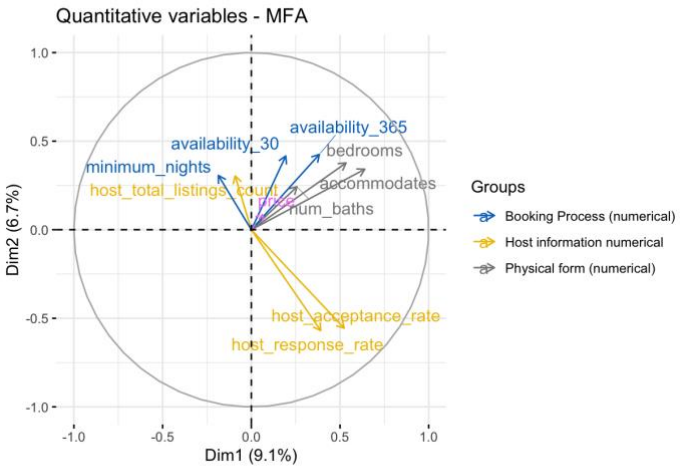


Image 72: PCA Variables Factor Map

We can also plot the individual variables of our MFA in a variable factor map that we already saw in the PCA. We can observe that host_acceptance_rate, availability, bedrooms and accommodates contribute a lot to Dim1.
Furthermore, we can deduce that the variables host_response_rate and host_acceptance_rate as well as the availability are contributing a lot to Dim2, and we can also see that they are negatively correlated with the host_acceptance_rate.

We are now continuing by exploring dimension 3 and 4.

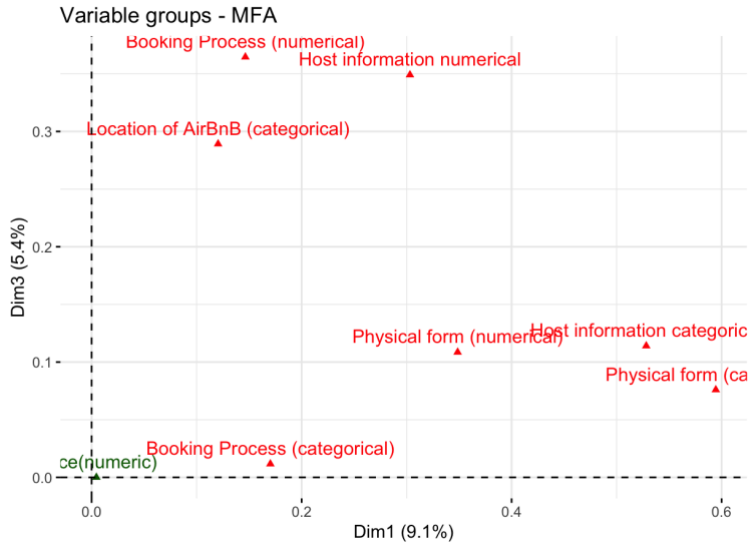


Image 73: Variable groups MFA

We can see that our Dim3 is mainly described by the numerical variables for the booking process and the host information, which was already the case for Dim2. However, this time the location of the AirBnB also plays a significant role.

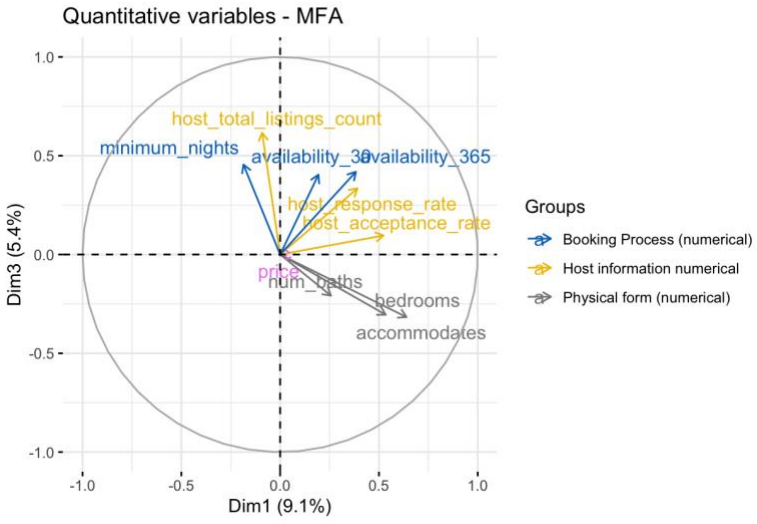


Image 74: PCA Variables Factor Map

In terms of quantitative variables, we can see that the variable host_total_listings_count is contributing a lot to dimension 3. Moreover, all variables from the group Booking Process have quite a significant contribution.

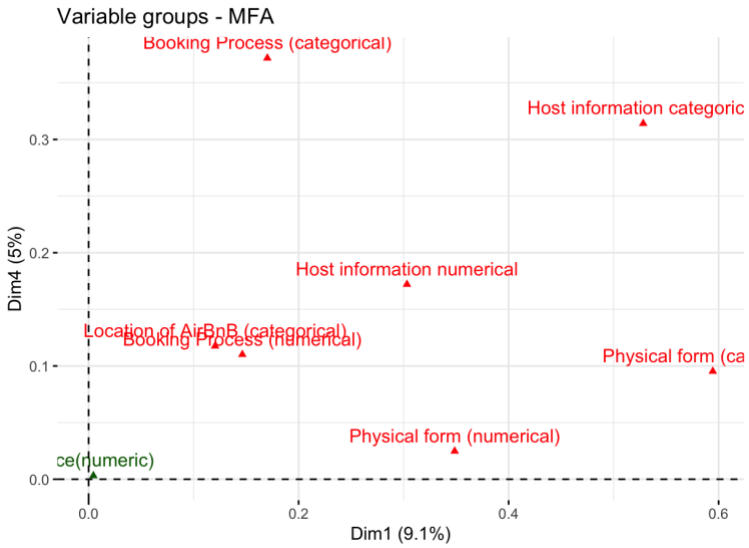


Image 75: Variable groups MFA

Considering dimension 4, we can see that the groups Booking process and Host information are the ones that contribute to it the most. However, this time it is the categorical groups of those variables.

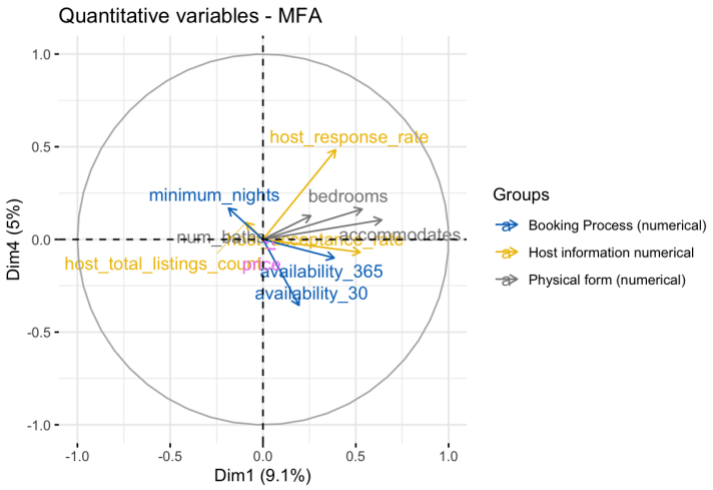


Image 76: PCA Variables Factor Map

For our Dim4, host_response_rate is actually the most contributing variable. Availability is the only other variable that has a significant impact on this dimension.

Finally, we decided to do one single plot of Dim1 and Dim2 plotting all individuals. As it is rather hard to make conclusions from this graph due to the low variability explained overall, we kept it to those two dimensions. Individuals with similar profiles (so AirBnBs with similar profiles) are close to each other in this plot We have a group of points in the first quadrant that are good considering our Dim2, which represented the ease of the booking process, but not so good when it comes to the physical form of the AirBnB. This area could for example be an interesting starting point when going into further analysis.

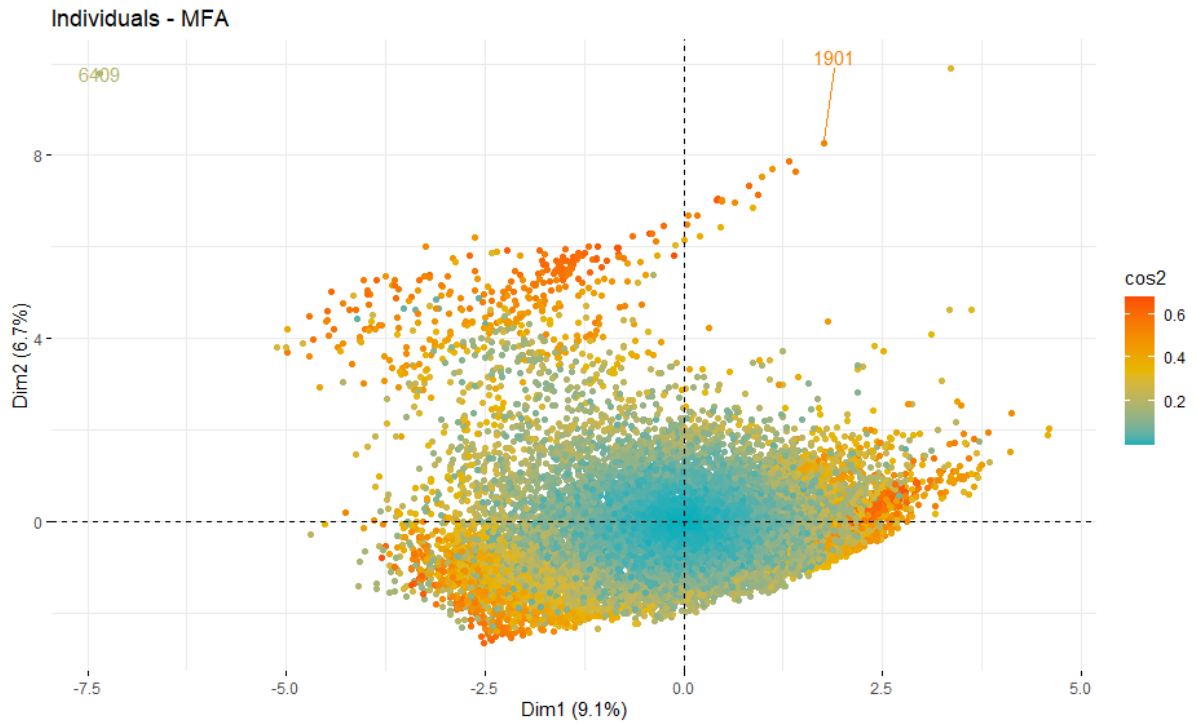


Image 77: Individuals Plot MFA

12. Association rules mining analysis

Association rules are "if-then" statements that demonstrate the connections between facts. They are typically performed for market basket analysis; for example, if a client purchases butter, he is most likely to buy bread (the traditional goal of an Association rule). However, in our project we would like to utilise them in an innovative way - to discover indications of Airbnb pricing or rating listings in Barcelona, as well as linkages between accommodation characteristics such as location, amenities, host characteristics, and so forth. The two best-known fundamental algorithms for mining frequent sets of items in a series of transactions are Apriori and Eclat. In this section, we elaborated the implementations and applications of these two techniques in our data set.

We proceeded by investigating links between listings for categorical variables. Given that we focused primarily on the reviews given to the listings and their prices throughout this paper, we followed the same approach and integrated our two 'target variables' into this study as well. Thus, we converted the aforementioned variables into categories after computing and analysing their distributions in this manner:

Review_scores_rating- Since most of review scores are close to 5, we decided to apply the following three ranges:

Low: for reviews lower than 4
Medium for reviews that are higher than 4, but lower than 4.6
High: for reviews higher than 4.6

Price- Since most of the prices are lower than 500, we decided to split the ranges accordingly:

Low: Prices below 100 euros/ night
Medium Low: Prices between 100 and 300/ night
Medium High: Prices between 300 and 500/night
High: Prices higher than 500/night

We excluded the "*first_review*" and "*last review*" variables since they would not contribute with any new information about the aforementioned two. Consequently, our Association analysis now includes a total of 23 categorical variables. As depicted in the barplot below, we can see that there are quite a few modalities with a high item frequency rate, a behaviour we had already seen in the univariate analysis. Therefore, we considered performing our next steps focusing on "lift" for Apriori since it would bring more meaningfulness of the rules (Note: it can be seen that all the graphs below are ordered by lift).

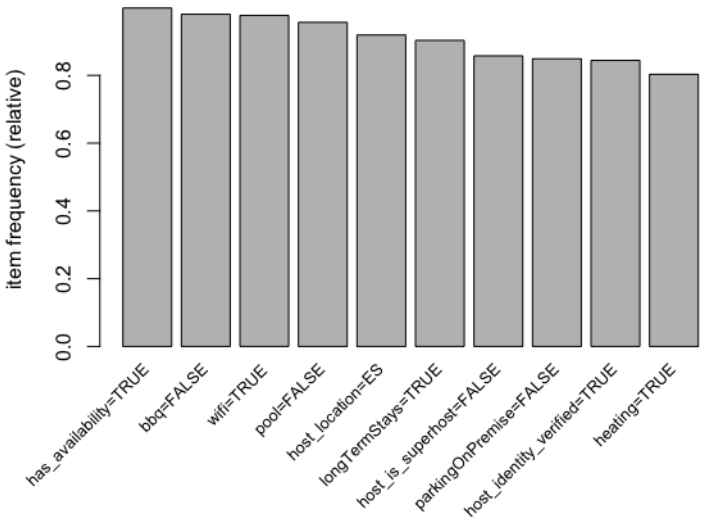


Image 78: Frequency Plot on single modalities

In the graph below, there are the most important length 2 20 rules for Apriori, in terms of lift, that have a minimum of confidence=0.5 and support=0.1. As we can see, most of the relations encompass typical modalities present in “private room” listings, which are the majority part in our Airbnb dataset. The lack of amenities, a shared bathroom or the exemption in license are all typical characteristics.

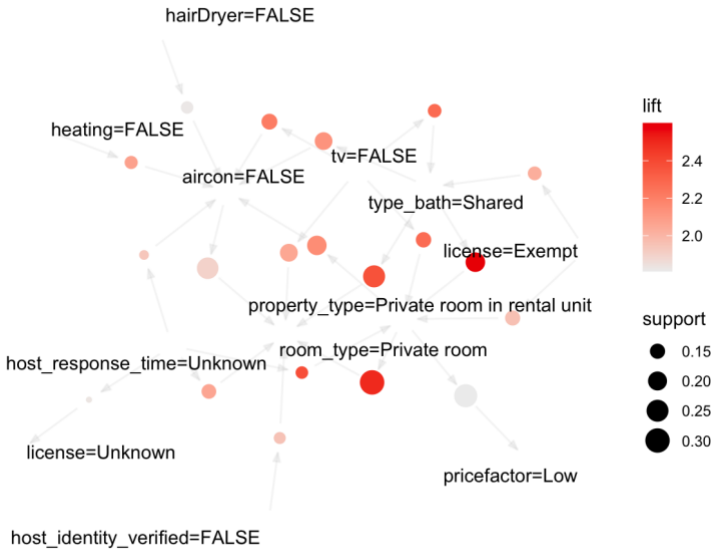


Image 79: First 20 rules of length 2 after performing Apriori, ordered by Lift

We have also continued our Rules analysis by maximising the number of items that are included in the rules to [3,5] (updated minimum length to 3 and maximum length to 5), while preserving the above mentioned support and confidence levels. Hence, ordered by lift, we observed that:

-If the host is not a superhost and the accommodation is a private room in a rental unit that lacks parking, barbecue, or outside area, it is most likely impossible to estimate the host's response time.

-If you have a private room as your room type and paid a low price for it, and you also lack various facilities such as a pool, air conditioning, barbecue, or parking, you most likely booked a private room in a rental unit, which is self explanatory.

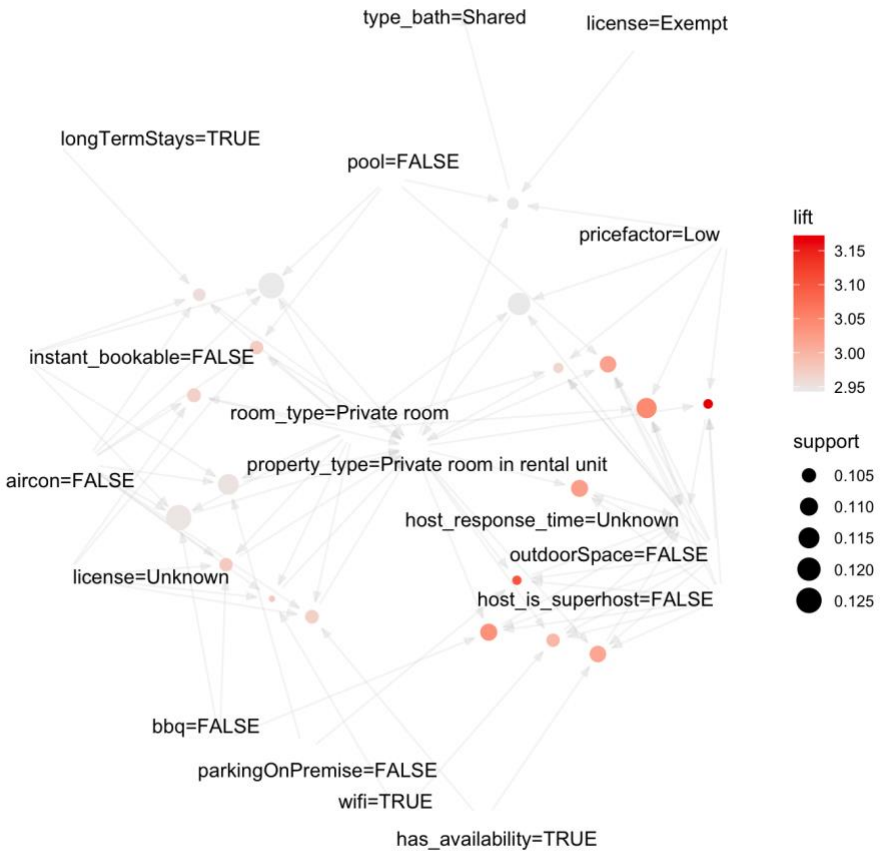


Image 80: First 20 rules of length 3 to 5 after performing Apriori, ordered by Lift

We, again, conducted the same research, but this time we concentrated on what might result in paying a cheaper amount for the rentals. That is, we filtered the rules to keep only the low prices on the right side of the associations. In summary, it appears that reserving a private room with a common bathroom and foregoing some facilities would result in a reduced nightly rate (Note: We concentrated on low pricing since, according to the price histogram, it is the most common pricing.). Also, the host living in Spain and good ratings seem to lead to a cheaper price.

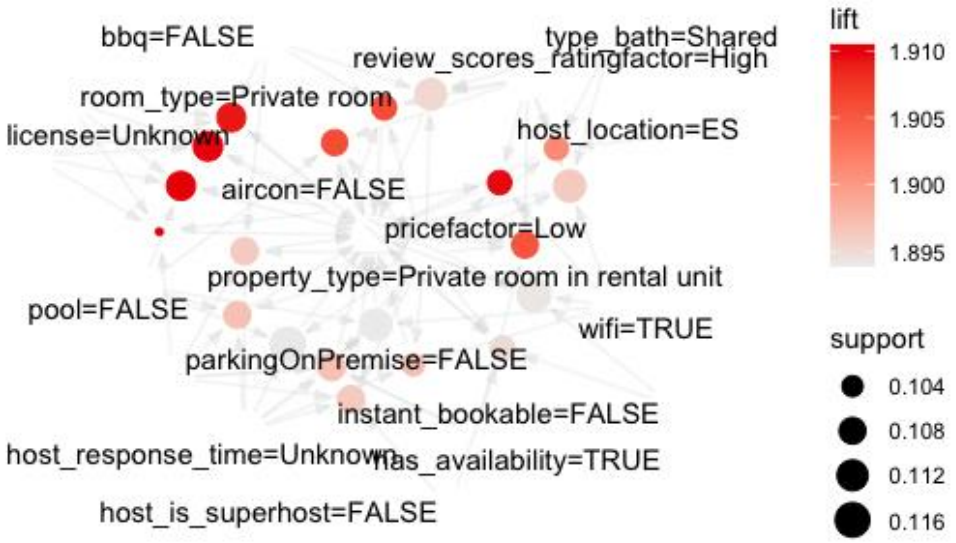


Image 81: First 20 rules of length 3 to 5 for low price after performing Apriori, ordered by Lift

Adopting the same rationale, we centered on what would result in a high rating score (the most common case, seen in boxplot analyses). As predicted, if you book an available room from a superhost (in ES) with multiple common facilities (hairdryer, internet, heater), you will most likely leave a favourable review regarding your Airbnb experience. Surprisingly, high-end facilities like a pool or a barbeque seem to have no effect on the high ratings in this graph.

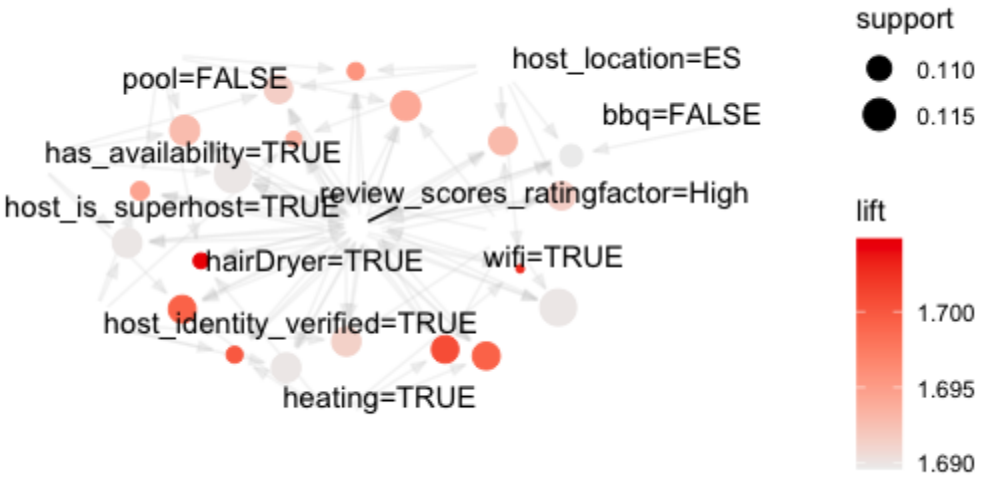


Image 82: First 20 rules of length 3 to 5 for high ratings after performing Apriori, ordered by Lift

Still searching for bigger length meaningful associations, we found an interesting rule for *ratingfactor=Low*. A listing with a host that is not a superhost, with an unknown license, parking not available and that permits long term stays is related with a low rating (or no rating, given the imputation method we applied). The lift is 1.92 and the support 0.11 with a confidence of 0.6.

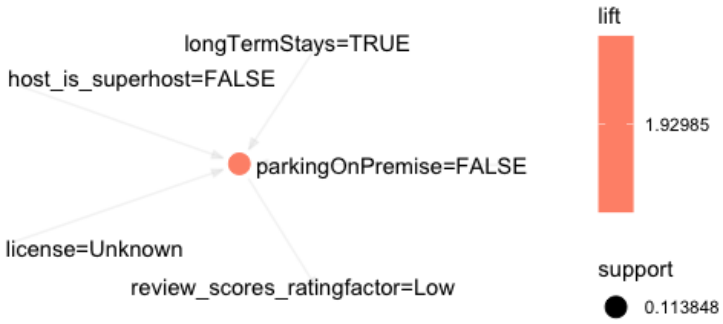


Image 83: First 20 rules of length 5 for low ratings after performing Apriori, ordered by Lift

Going back to analysing pairs (length 2 associations) of items, we took into consideration the highest rating factor score, i.e. *reviews_scores_ratingfactor=High* as the right side of the association, the table below indicates that :

-If the clients rent an entire rental unit
-If the listing is located in Ciutat Vella or Eixample (i.e. central)
-If the price associated with the listing is either low or high
-If the listing you rented has outdoor space
-If the host posting the listing is considered a superhost

Then the client is most likely going to leave a high review on your listing.
They seem reasonable since the neighbourhoods are considered good areas for tourism, an outdoor space is a very appreciated amenity, and paying a reasonable price for these would lead to a happy customer.

However, we discovered that even those renting over shared accommodation, such as a private room in a rental unit, that do not provide air conditioning or television are indeed satisfied with what they received. Furthermore, the license and instant bookable status have no influence on the scores.

Image 84: First 20 rules of length 2 for high scores after performing Apriori, ordered by Lift

We reduced the support to 0.001 and the confidence to 0.1 in order to check association rules for the *review_scores_ratingfactor=Low*, given the low percentage of low scores in the dataset. The graph below depicts all of the factors that might result in a negative review of an accommodation. As a result, a higher fee, a lack of WIFI, a slow response from the host, a lack of heating or a hairdryer will almost surely result in a negative review. Again, location is key because lodgings in Sant Andreu, Nou Barris and Sarria Sant Gervasi are most likely to receive a low rating. The first two correspond to poorer districts and far from the city centre.

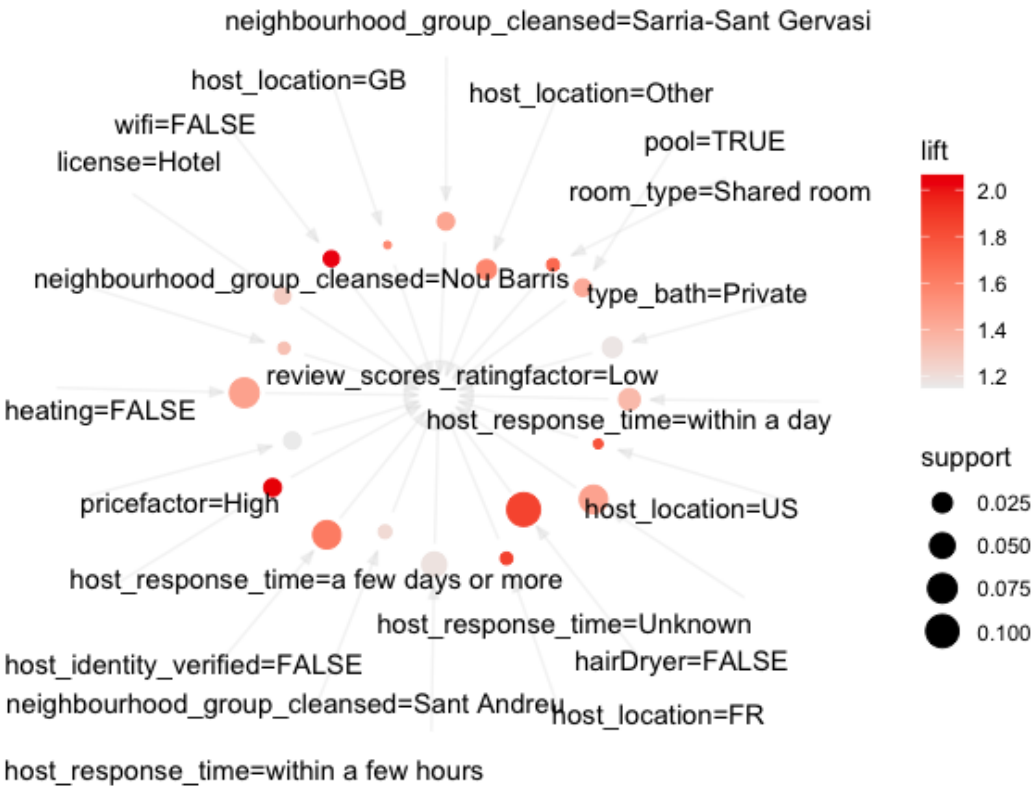


Image 85: First 20 rules of length 2 for low ratings after performing Apriori, ordered by Lift

When it comes to paying a lower price, i.e. *pricefactor=Low* we can see may have an impact on some of the characteristics of the accommodation, such as:

-If you lack air conditioning, TV, heating, hair dryer
-If you booked the room from an unverified host
-If the room the client booked is shared (i.e. private room)
-If the bathroom is shared
-If the reviews obtained are low

Then the client most likely paid a small amount of money for the accommodation.
However, an intriguing rule we noticed is that a room in 'Ciutat Vella' implies a lesser price, which in fact, given its central location, may not be expected.

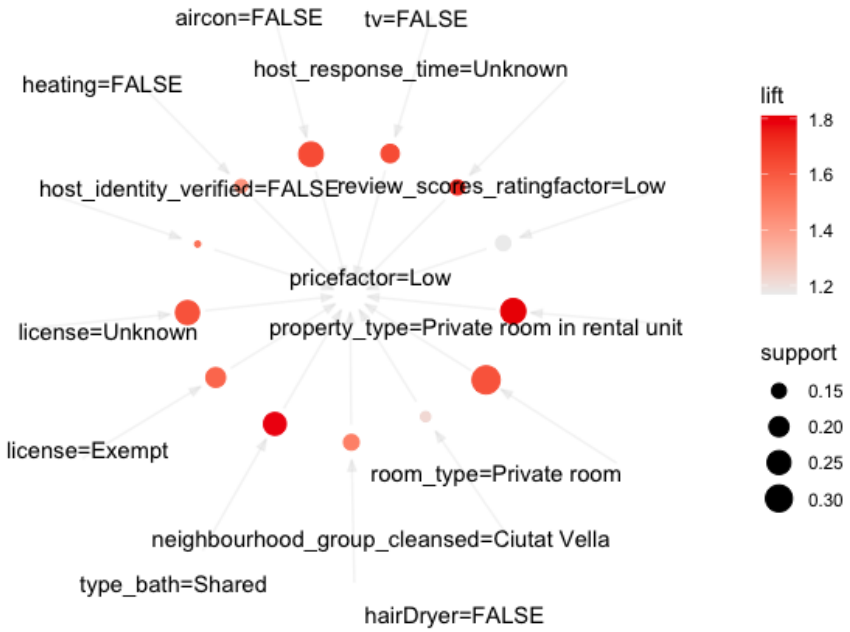


Image 86: Rules of length 2 for low prices after performing Apriori, ordered by Lift

In order to obtain association rules for the *pricefactor=High*, we have lowered the support to 0.001 and the confidence to 0.1 since there were not many entries with a higher price. Hence, we could conclude that staying in a hotel (which would imply that host response time to be higher than usual) or having barbecue or pool facilities increases the price significantly for the customer. Moreover, if the area the accommodation is located in is either Sarria-Sant Gervasi or Les Corts is most likely to increase the price per night of the room booked.

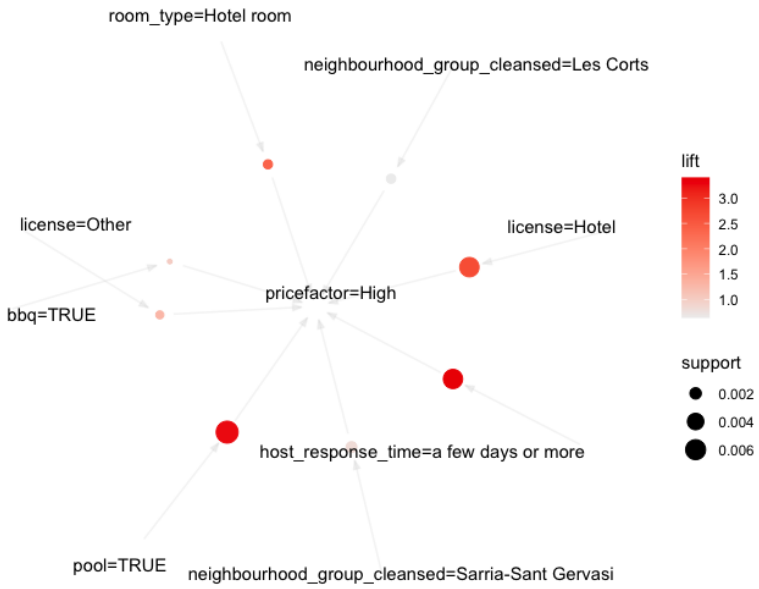


Image 87: Rules of length 2 for high prices after performing Apriori, ordered by Lift

Using the ECLAT algorithm, we obtained the top 20 most frequent itemsets with open length. The results can be seen in the image below. Most of them correspond to single or combined instances of the most common modalities in the dataset. Most of the listings are available and lack most of the possible amenities offered, given that most listings are low-cost. Having wifi is, however, rather common.

```
> inspect(top20)
     items                                                  support   count
[1]  {has_availability=TRUE}                                0.9987384 16624
[2]  {bbq=FALSE}                                            0.9803545 16318
[3]  {has_availability=TRUE, bbq=FALSE}                     0.9790928 16297
[4]  {wifi=TRUE}                                            0.9770502 16263
[5]  {has_availability=TRUE, wifi=TRUE}                     0.9760288 16246
[6]  {wifi=TRUE, bbq=FALSE}                                 0.9578853 15944
[7]  {has_availability=TRUE, wifi=TRUE, bbq=FALSE}          0.9568639 15927
[8]  {pool=FALSE}                                           0.9564434 15920
[9]  {has_availability=TRUE, pool=FALSE}                    0.9551817 15899
[10] {pool=FALSE, bbq=FALSE}                                0.9395614 15639
[11] {has_availability=TRUE, pool=FALSE, bbq=FALSE}         0.9382998 15618
[12] {wifi=TRUE, pool=FALSE}                                0.9342746 15551
[13] {has_availability=TRUE, wifi=TRUE, pool=FALSE}         0.9332532 15534
[14] {host_location=ES}                                     0.9191349 15299
[15] {host_location=ES, has_availability=TRUE}              0.9182938 15285
[16] {wifi=TRUE, pool=FALSE, bbq=FALSE}                     0.9177531 15276
[17] {has_availability=TRUE, wifi=TRUE, pool=FALSE, bbq=FALSE} 0.9167318 15259
[18] {longTermStays=TRUE}                                   0.9032142 15034
[19] {has_availability=TRUE, longTermStays=TRUE}            0.9020727 15015
[20] {host_location=ES, bbq=FALSE}                          0.9011715 15000
```

Also using the ECLAT algorithm, we plotted the 20 rules with most support with restricted length 2. Following the tendency seen in the frequent itemsets, frequent associations happen between lack of some amenities and basic properties of most hosts and listings.

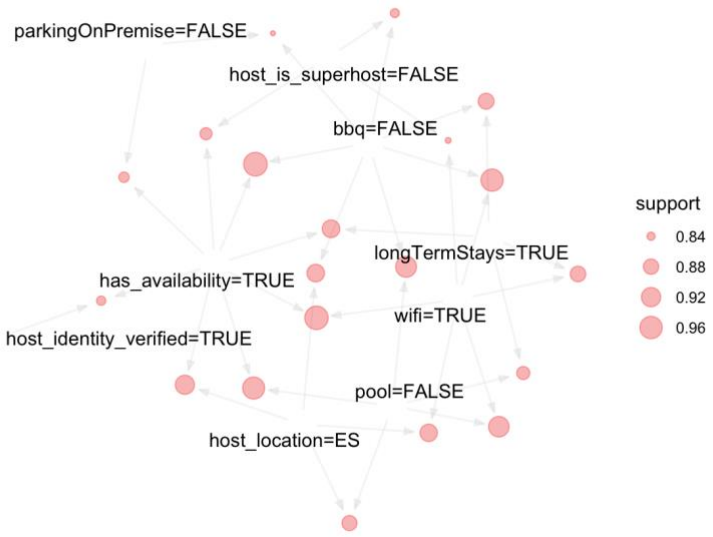


Image 89: First 20 frequent itemsets of length 2, after performing ECLAT

Performing rule induction based on the rules obtained with ECLAT, we outputted the plot with the top 20 rules by lift, respecting a minimum confidence of 0.5. Results are similar to the ones obtained with Apriori, which does not mean equal. The plot is very “Private room in rental unit” feature-based.

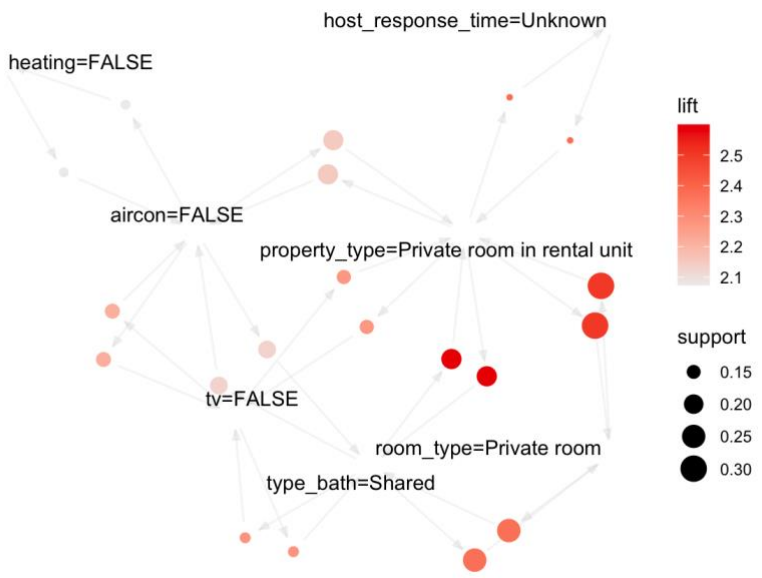
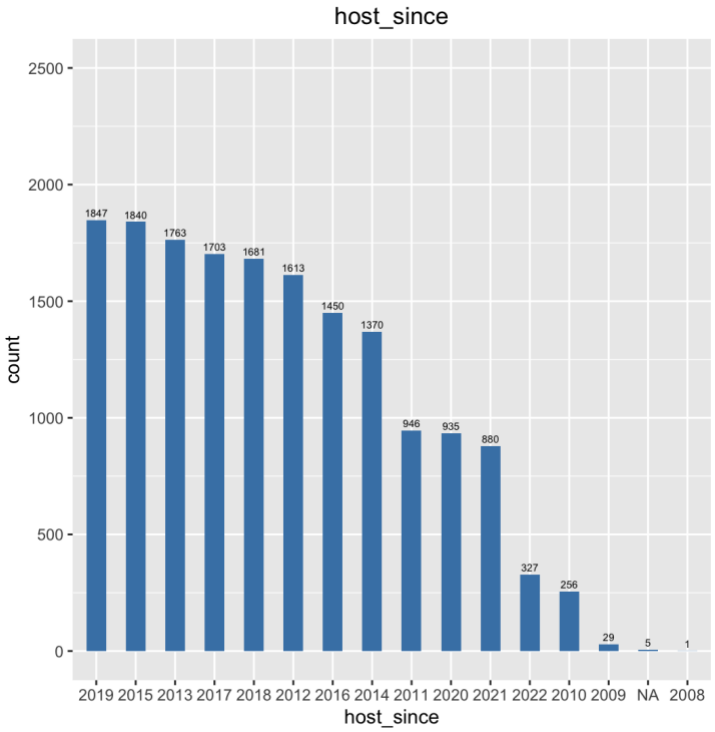


Image 90: First 20 rules, after performing rule induction ordered by lift

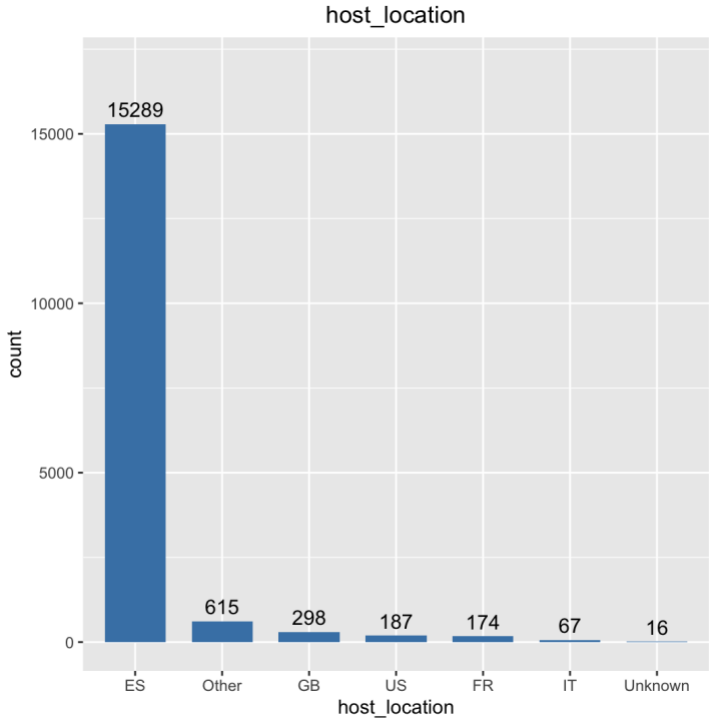
We believe that conducting the association rules algorithms on our data mostly led us to expected results. Both the hosts who placed these listings and the people renting and rating them appear to have relied on common sense. Since it is evident that a person booking a room in which you share the toilet and kitchen that lacks most facilities will most likely be charged a small fee, a person booking a hotel with a pool will be charged more. Furthermore, the reviewers appear to have taken a fair "what you pay is what you get" mentality, since even private rooms in rental apartments with a common toilet that do not have many facilities are scored well. Furthermore, it is plausible to deduce that the neighbourhood has a significant impact on both the ratings and the pricings of the listings.

ANNEX

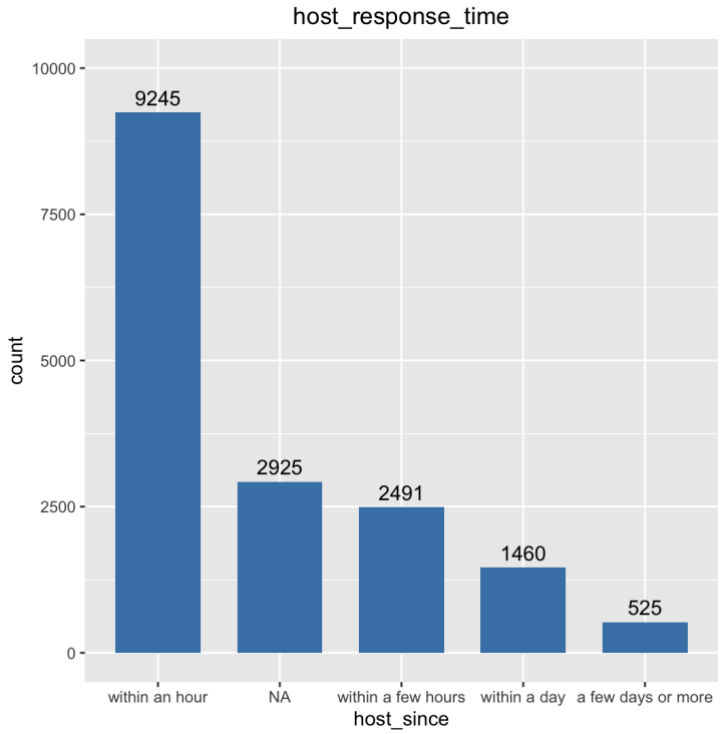
UNIVARIATE ANALYSIS



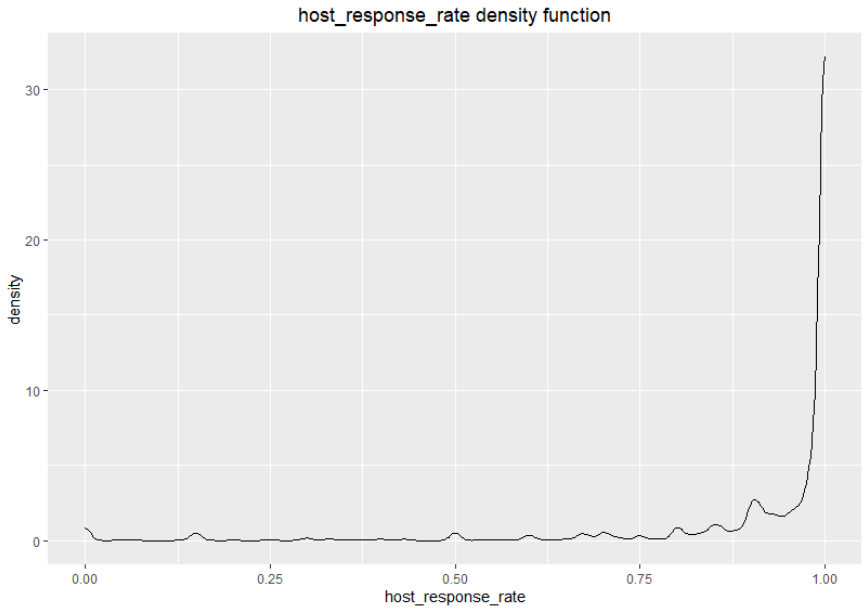
Barplot host_since



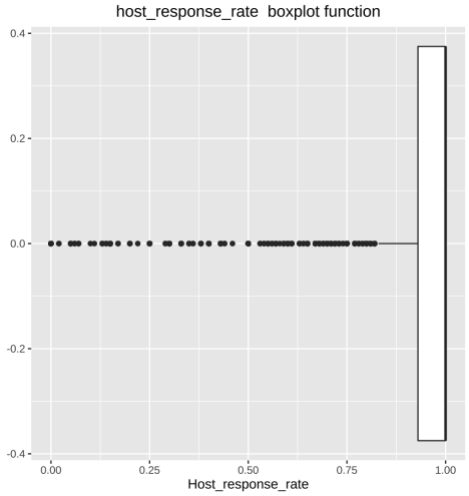
Barplot host_location



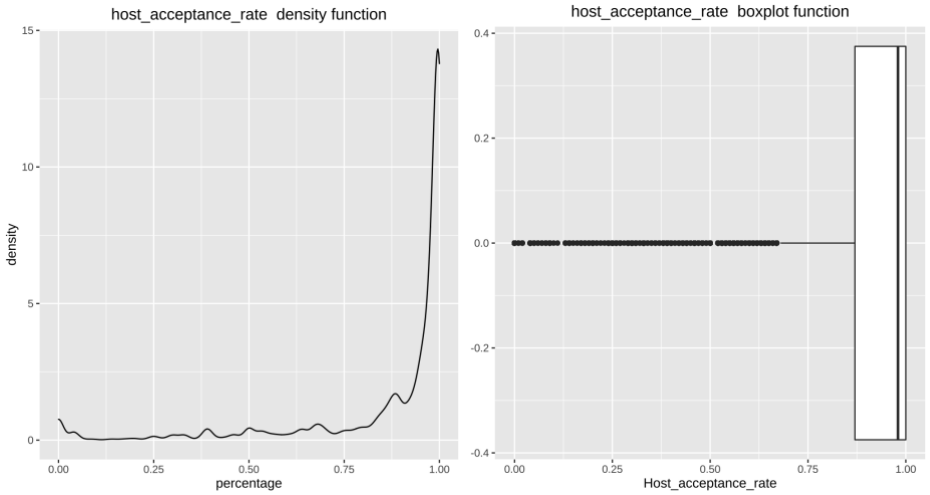
Barplot host_response_time



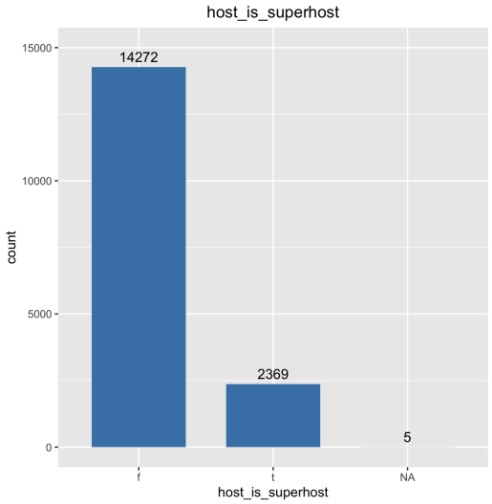
Density plot of host_response_rate



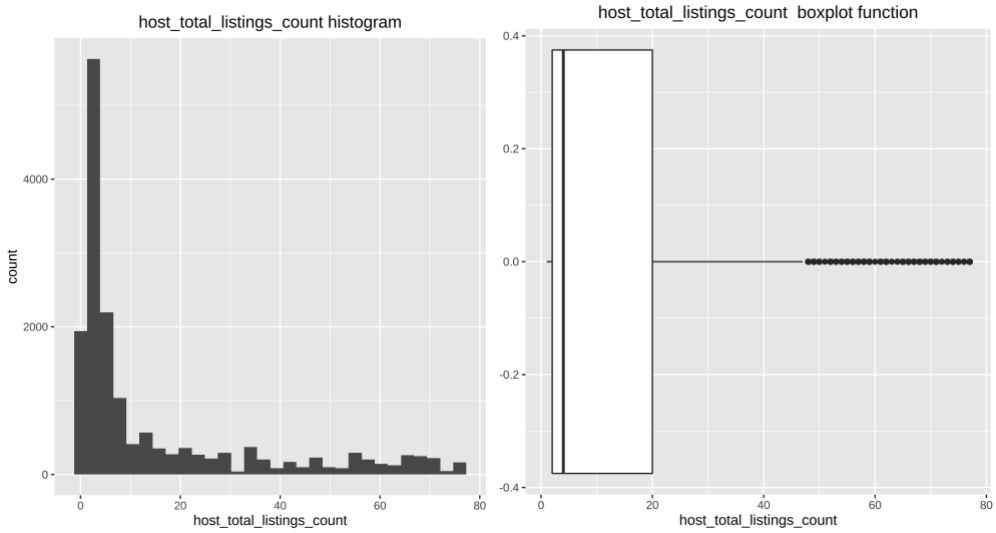
Boxplot host_response_time



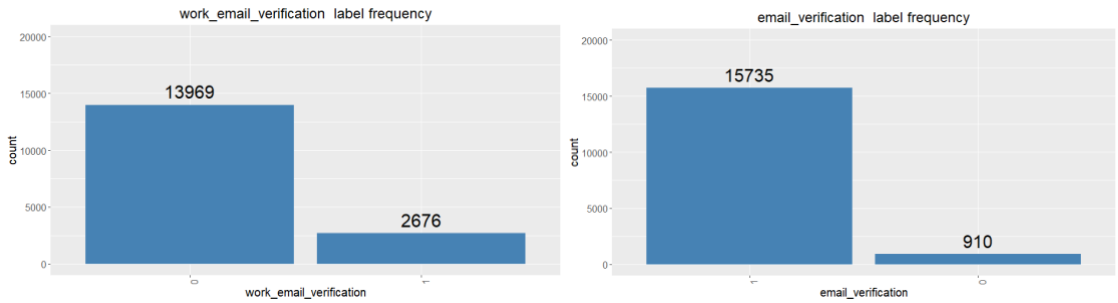
Density and boxplot of variable host_acceptance_rate



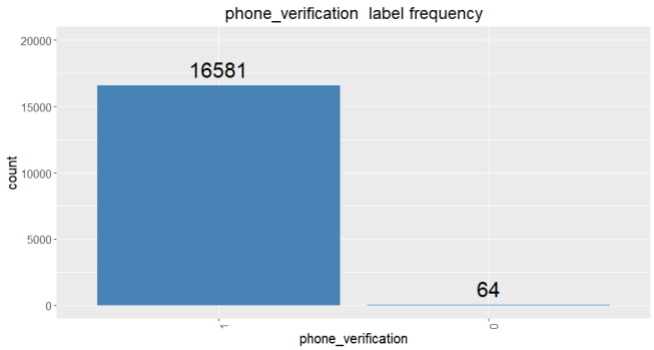
Barplot host_is_superhost



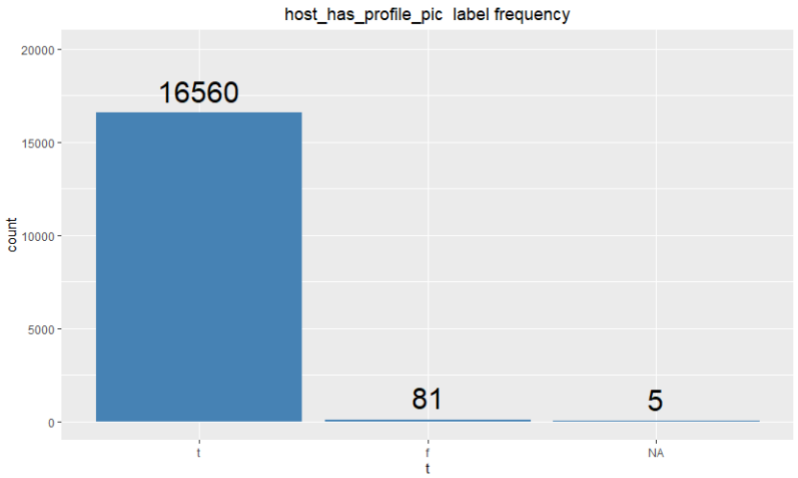
Histogram and boxplot of variable host_total_listings_count



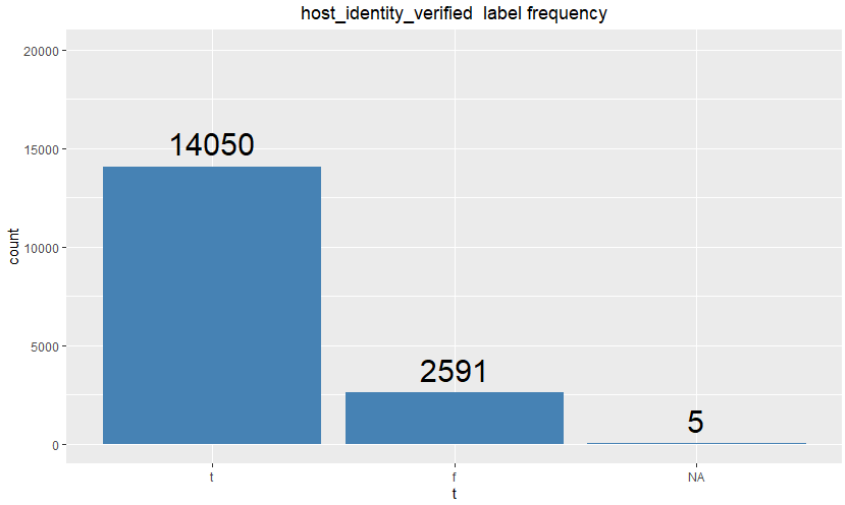
Barplots of work_email_verification and email_verification



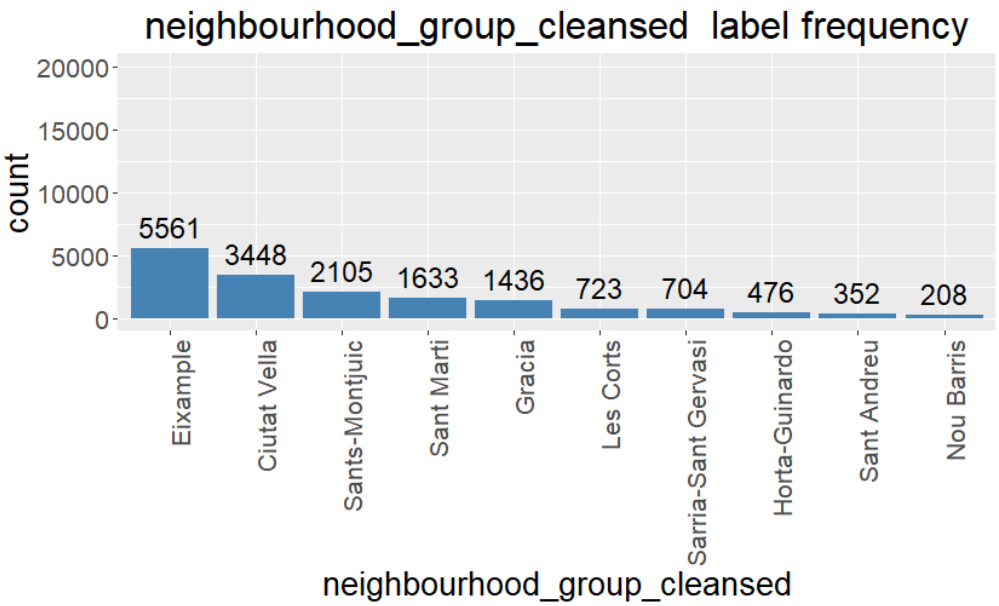
Barplot of phone_verification



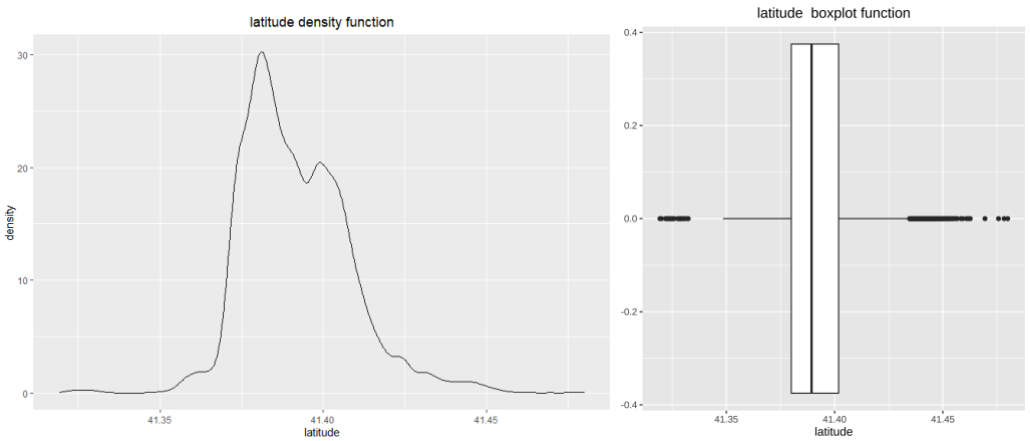
Barplot of variable host_has_profile_pic



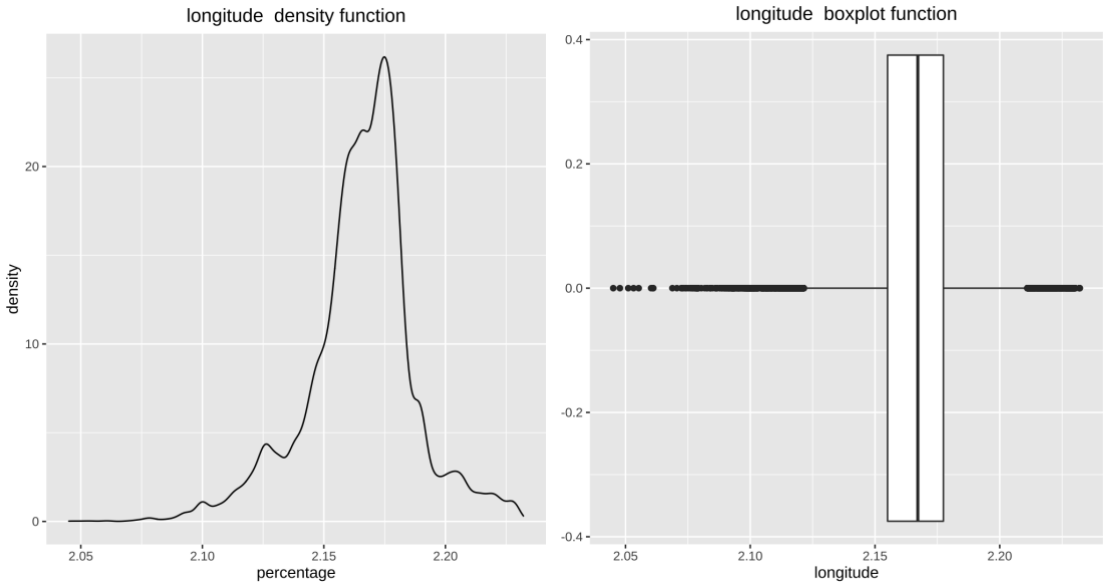
Barplot host_identity_verified



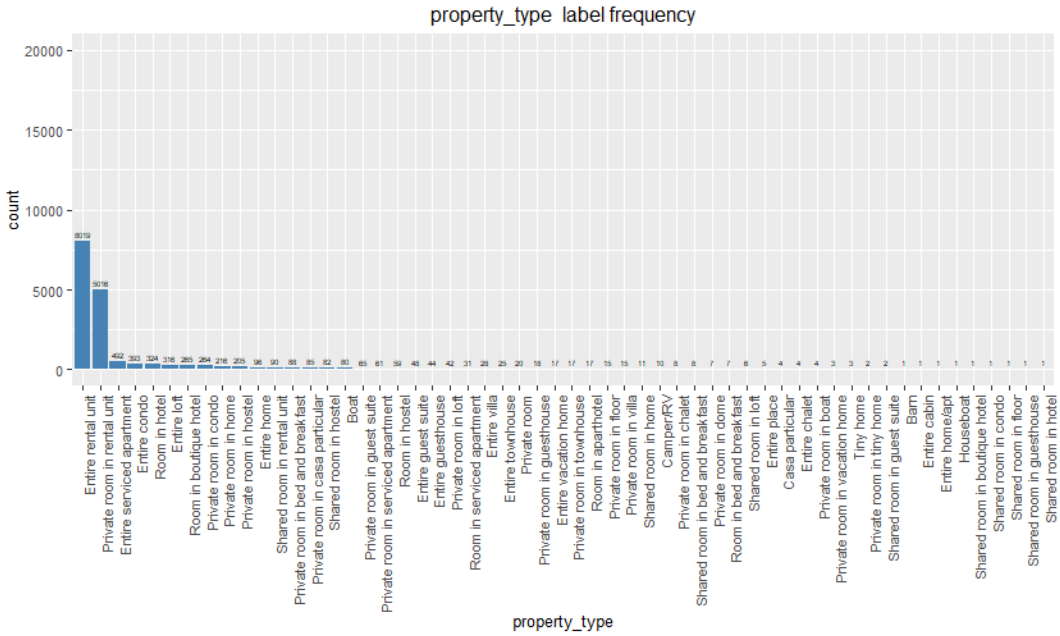
Barplot of neighbourhood_group_cleansed



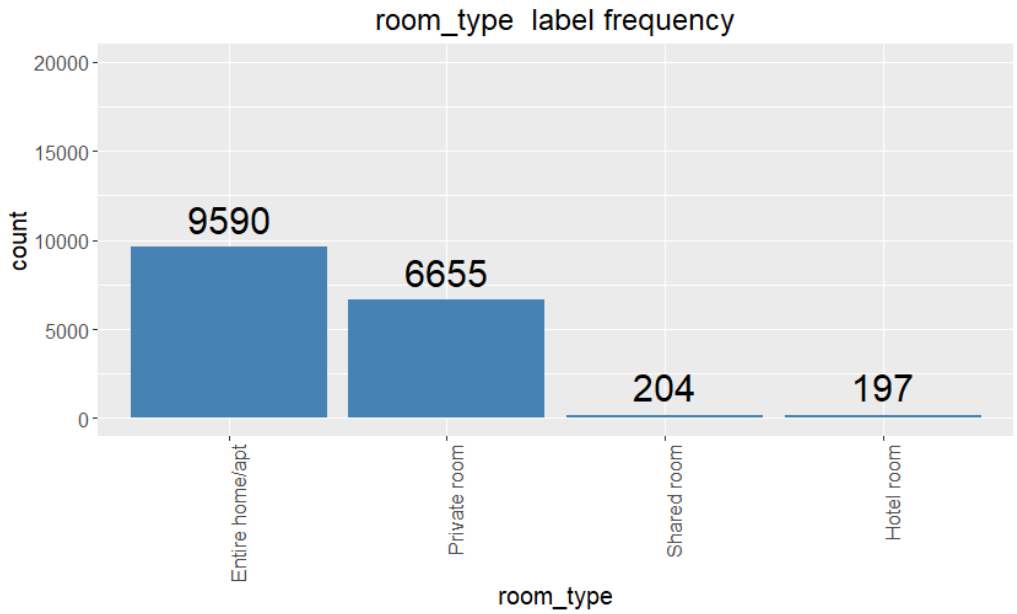
Density plot and boxplot of



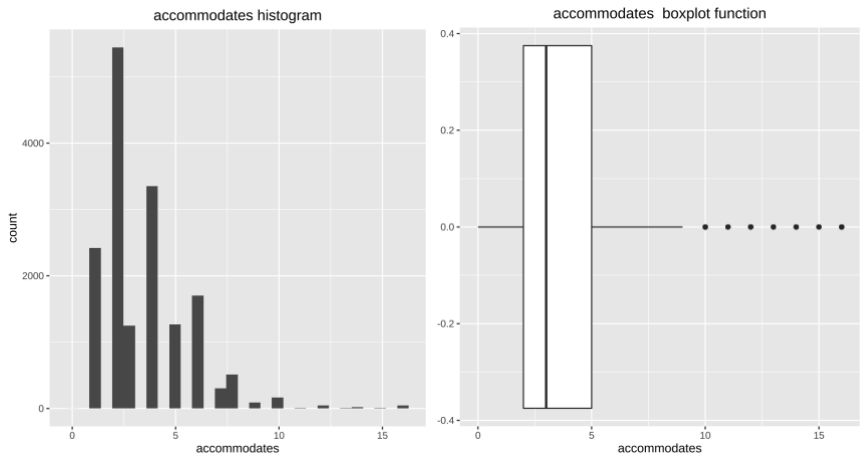
Density plot and boxplot of longitude



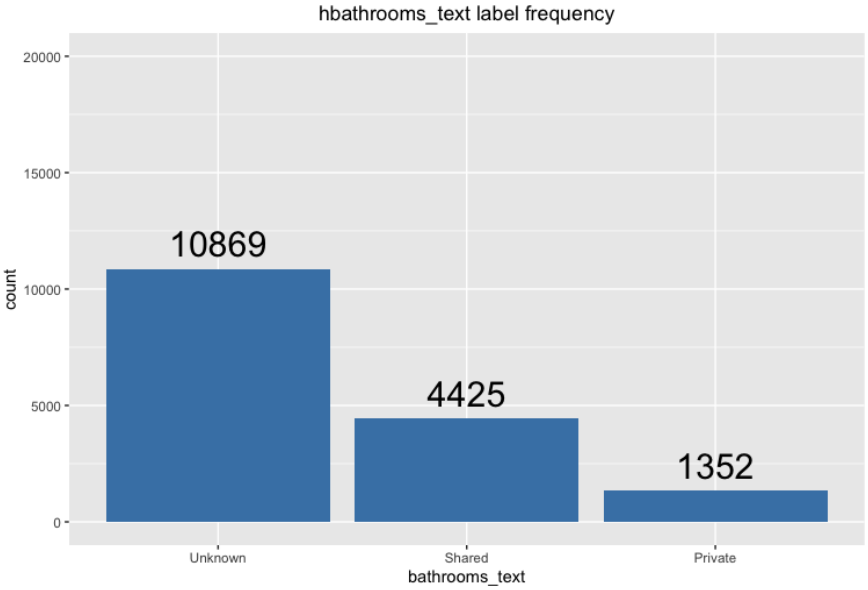
Barplot of property_type



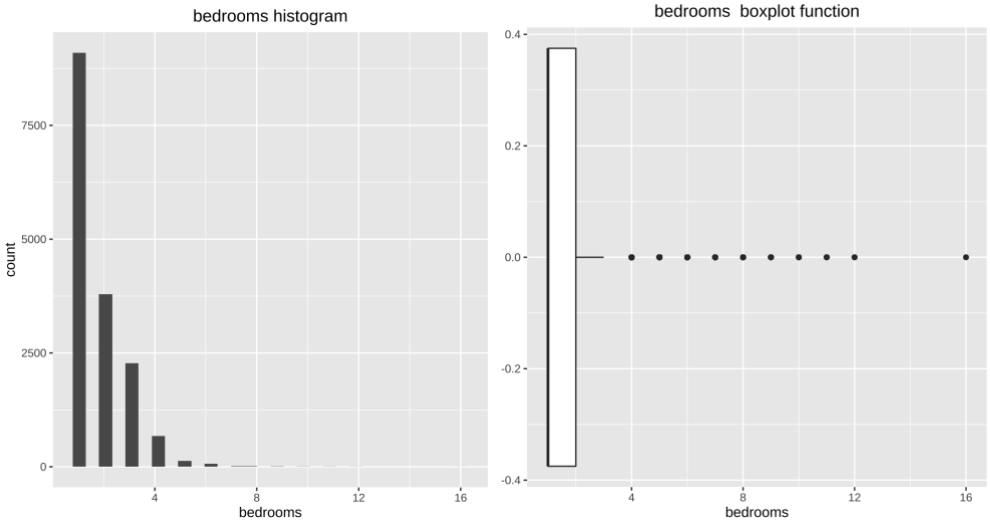
Barplot of room_type



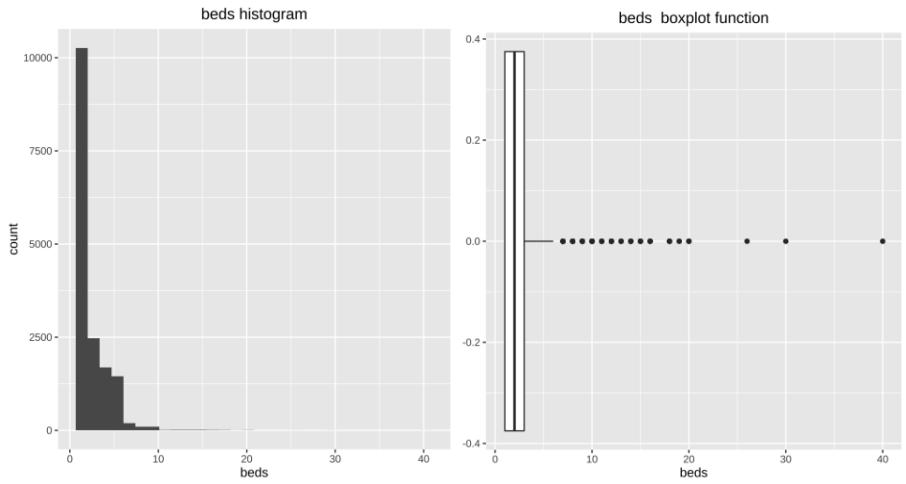
Histogram and boxplot of accomodates



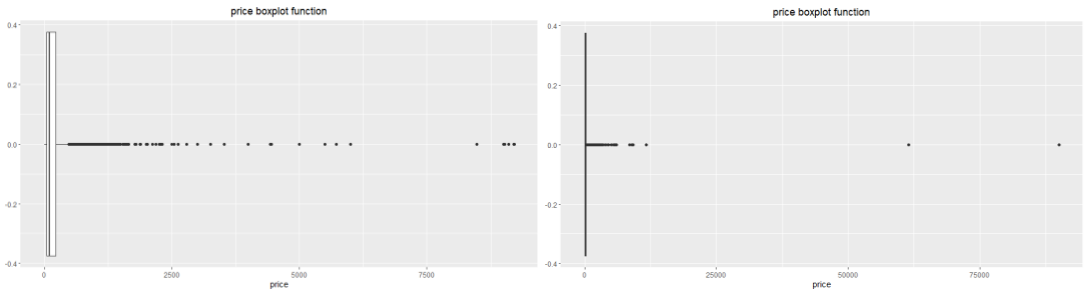
Barplot of bathrooms



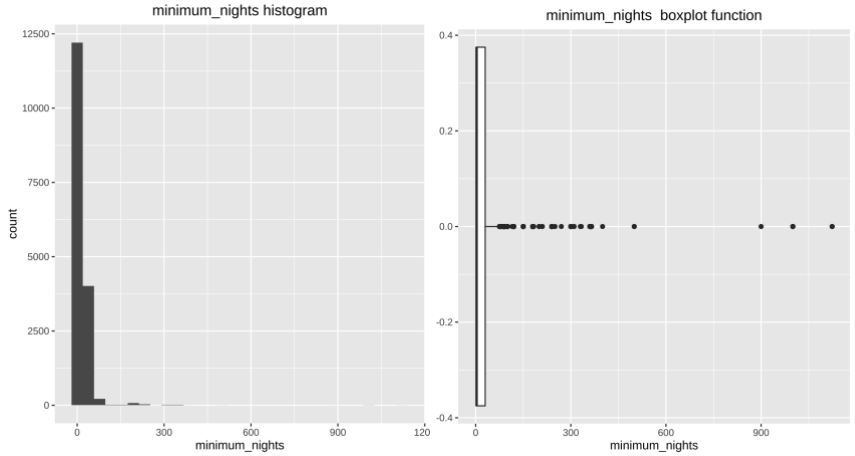
Histogram and boxplot of bedrooms



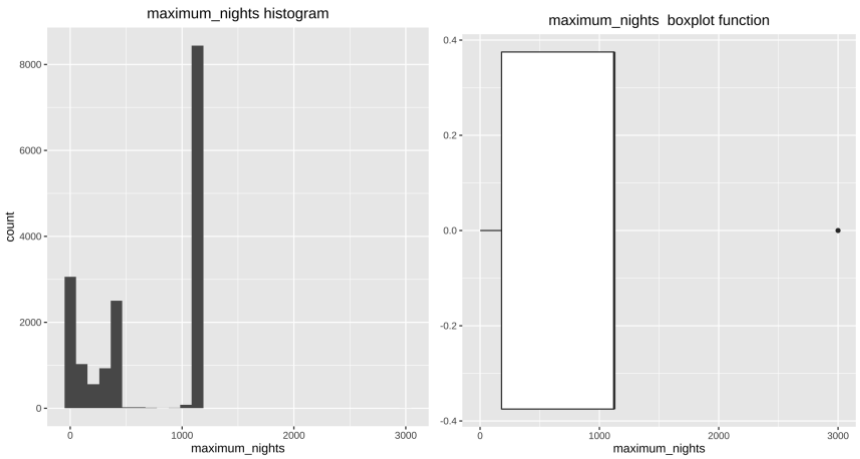
Histogram and boxplot of beds



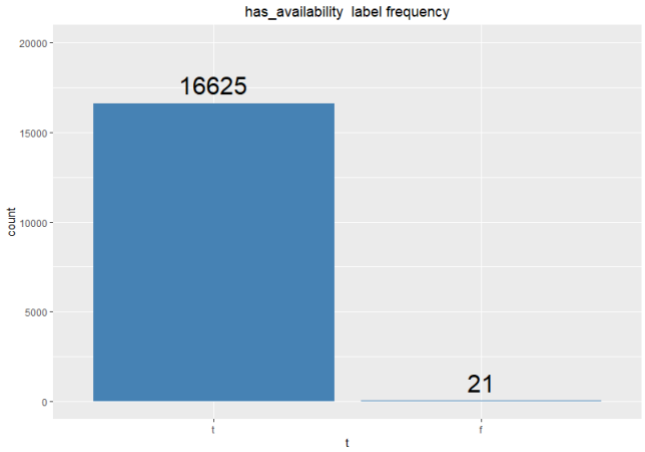
Boxplot of price



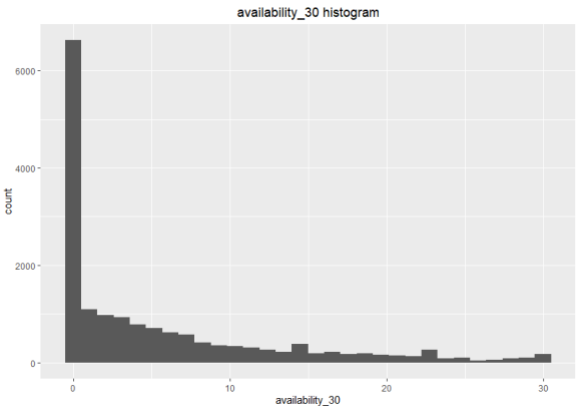
Histogram and boxplot of minimum_nights



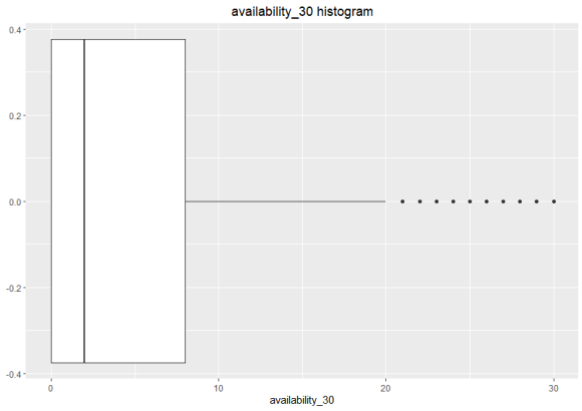
Histogram and boxplot of maximum_nights



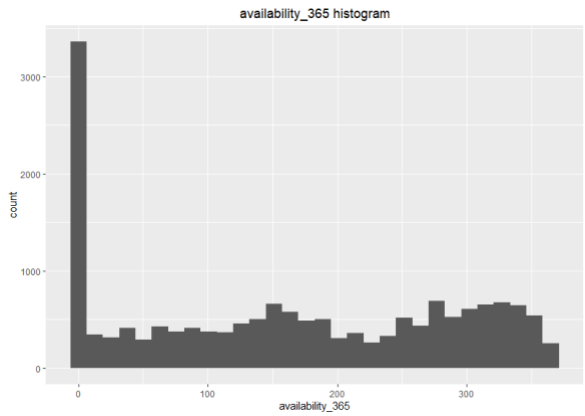
Barplot of has_availability



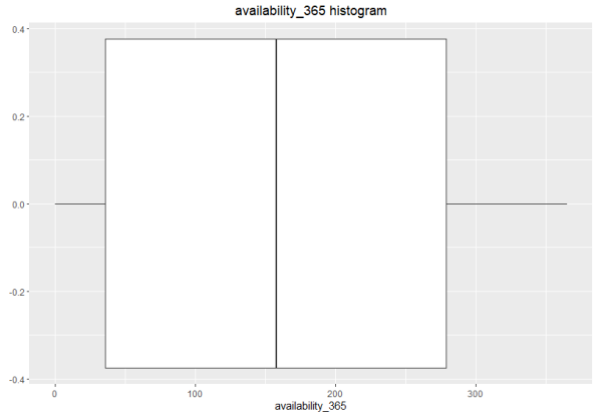
Availability_30 Histogram



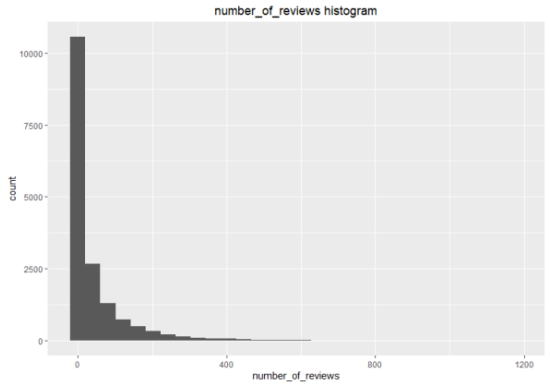
Availability_30 boxplot



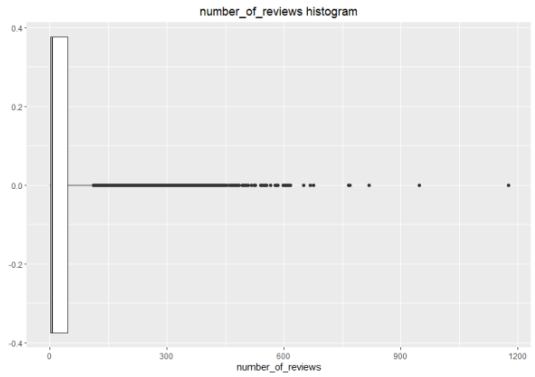
Availability_365 Histogram



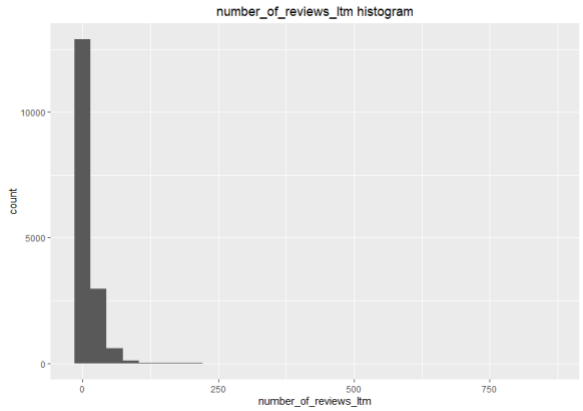
Availability_365 boxplot



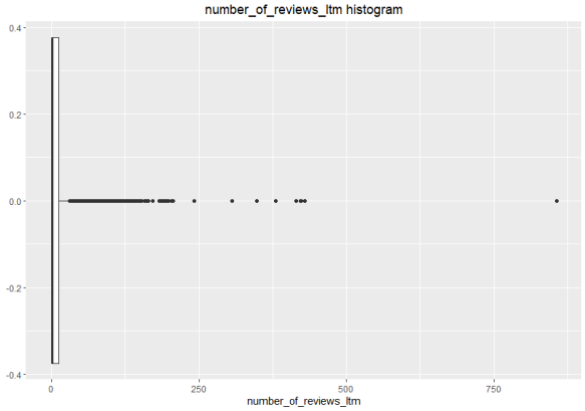
Number Reviews Histogram



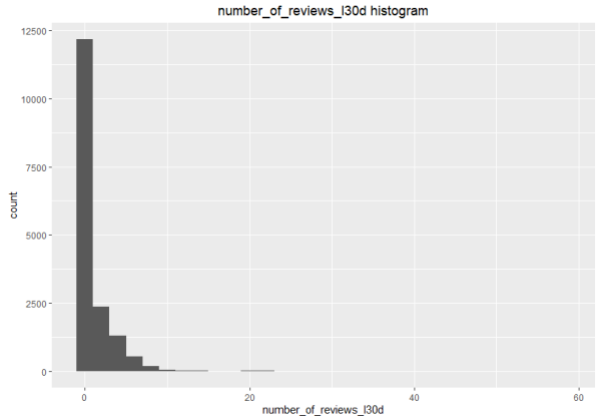
Number Reviews boxplot



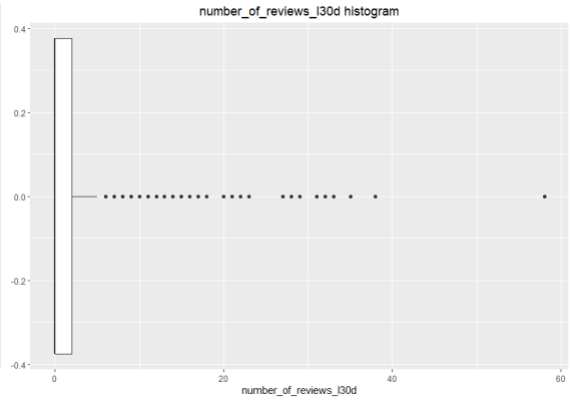
Number Reviews LTM Histogram



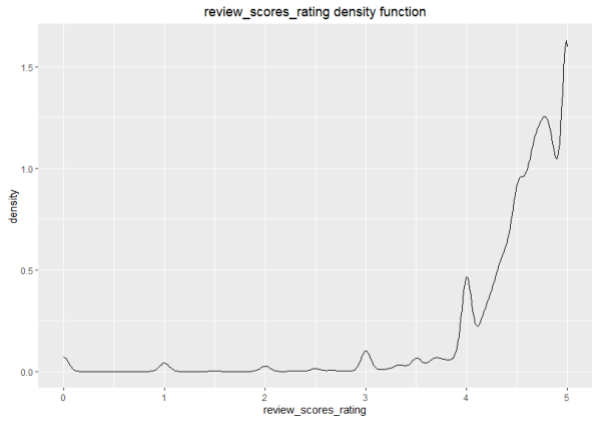
Number Reviews LTM Boxplot



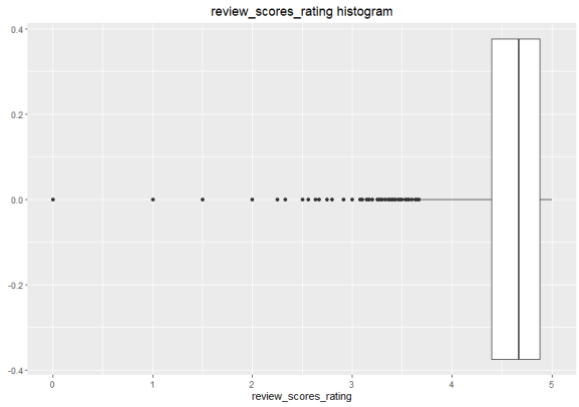
Number Reviews LTM L30D Histogram



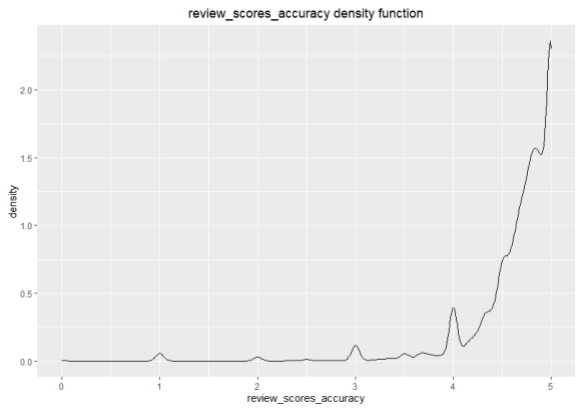
Number Reviews LTM L30D Boxplot



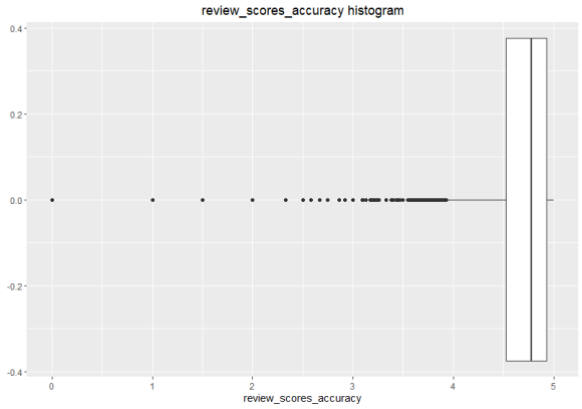
review_scores_rating Density



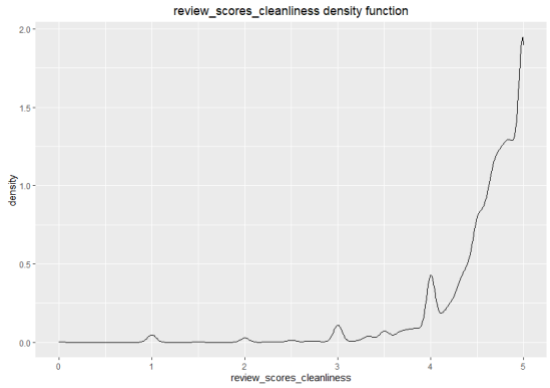
review_scores_rating Boxplot



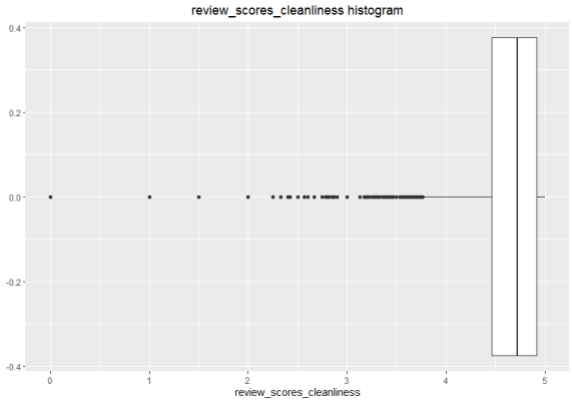
review_scores_accuracy Density



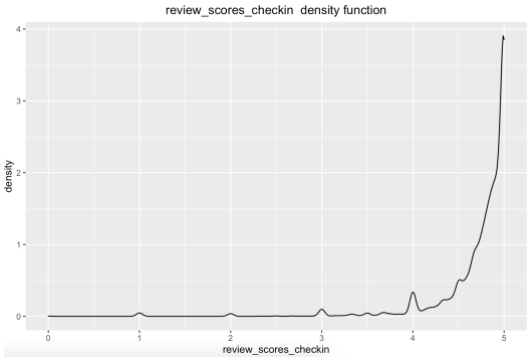
review_scores_accuracy Boxplot



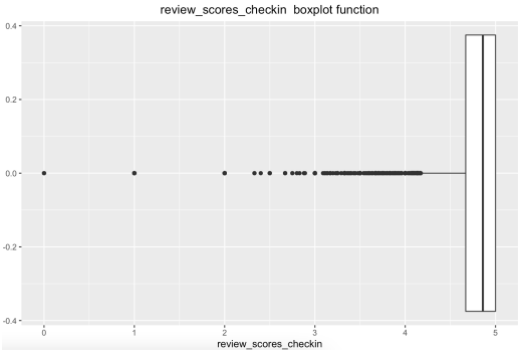
review_scores_cleanliness Density



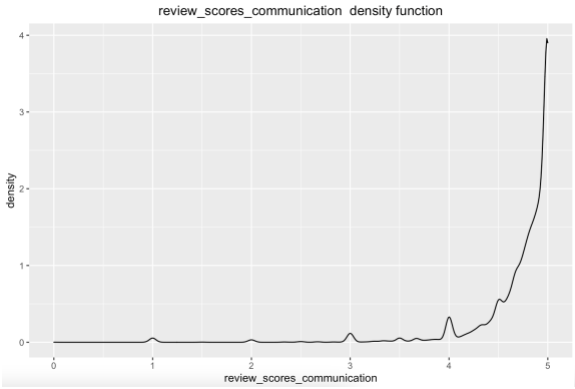
review_scores_cleanliness Boxplot



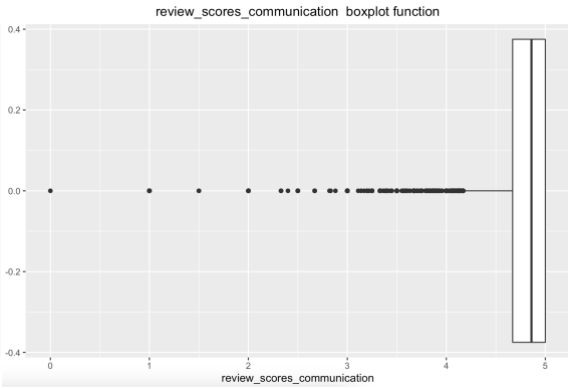
review_scores_checkin Density



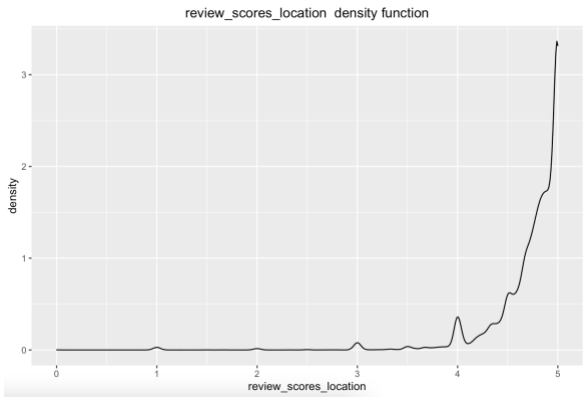
review_scores_checkin Boxplot



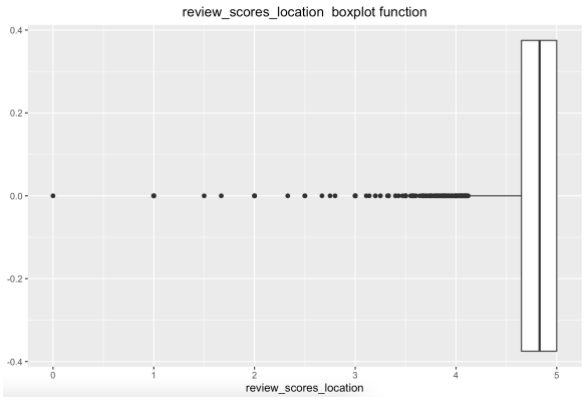
review_scores_communication Density



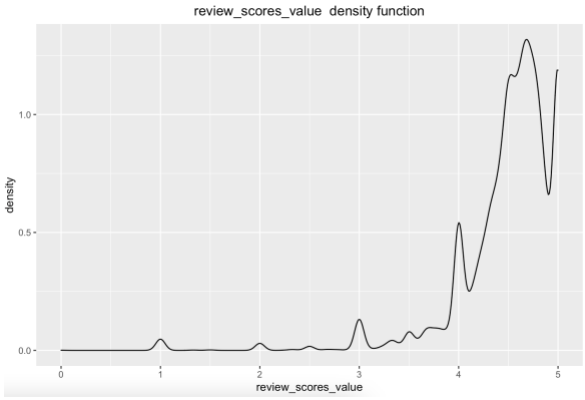
review_scores_communication Boxplot



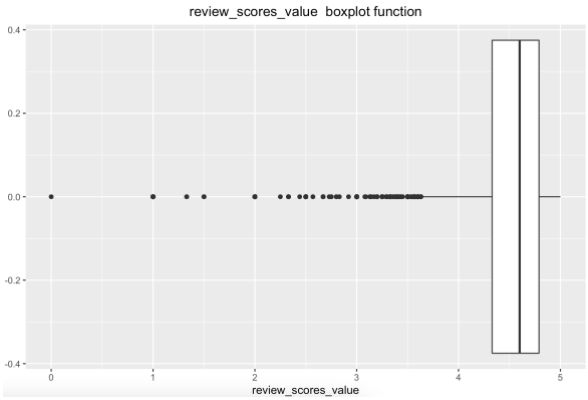
review_scores_location Density



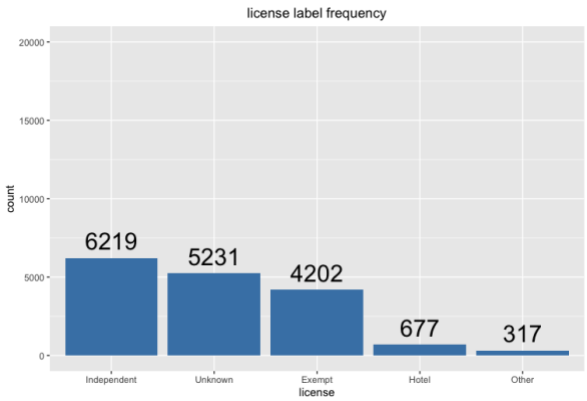
review_scores_location Boxplot



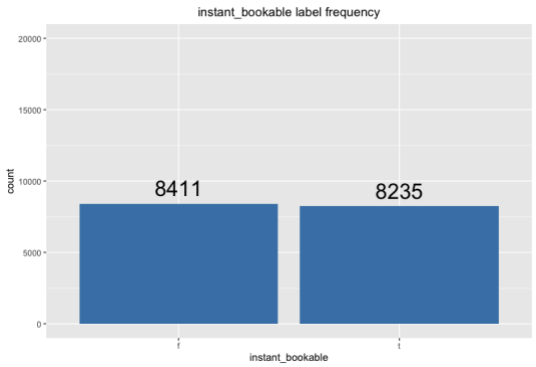
review_scores_value Density



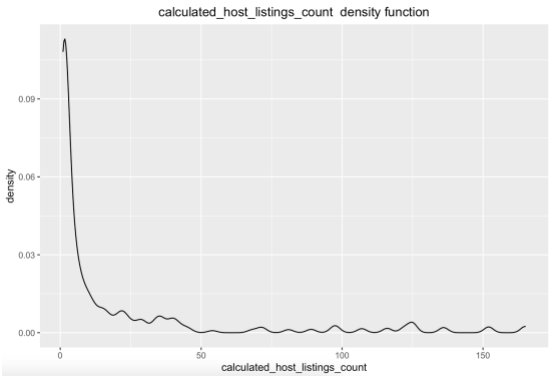
review_scores_value Boxplot



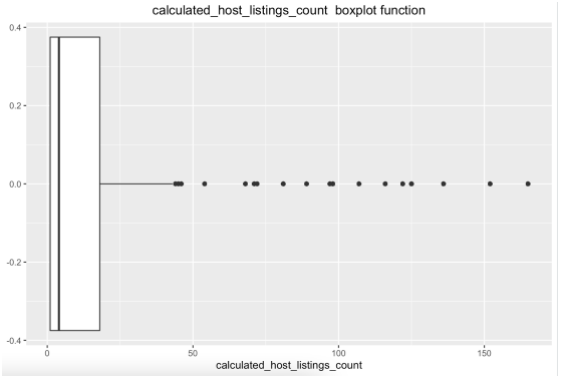
Barplot of license type



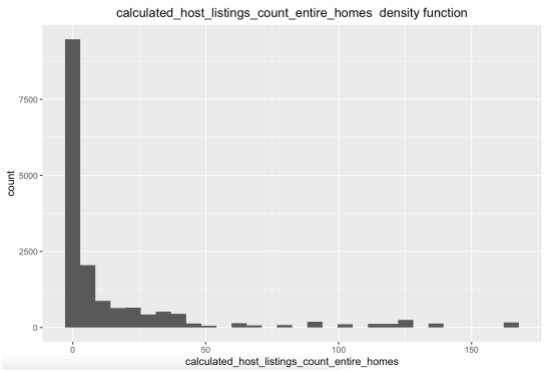
Barplot of variable instant_bookable



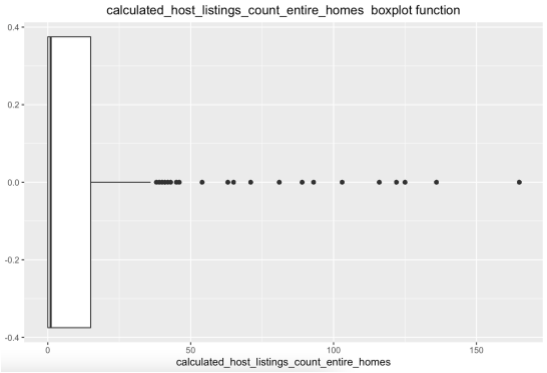
calculated_host_listings_count Density



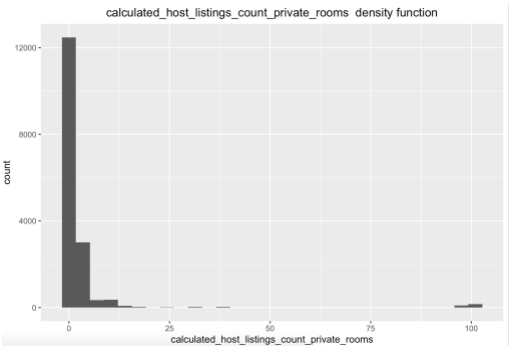
calculated_host_listings_count Boxplot



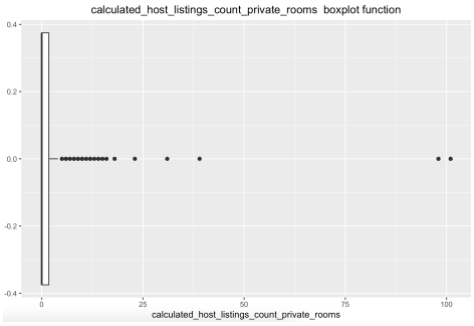
calculated_host_listings_count_entire_home Histogram



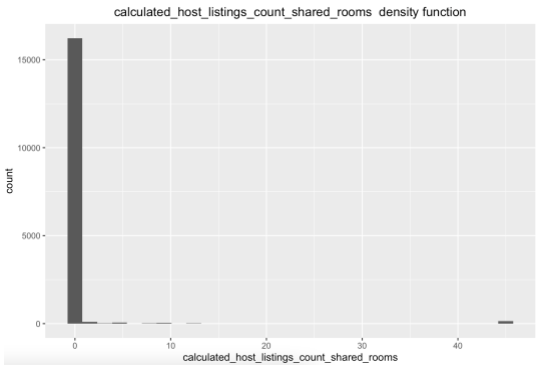
calculated_host_listings_count_entire_home Boxplot



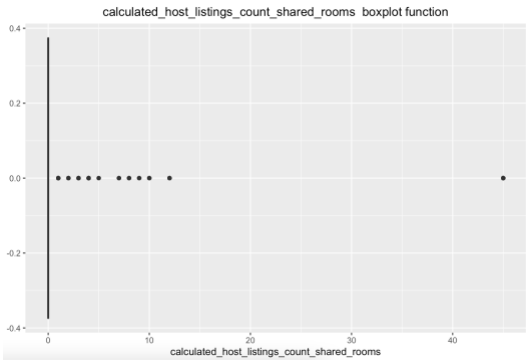
calculated_host_listings_count_private_home Histogram



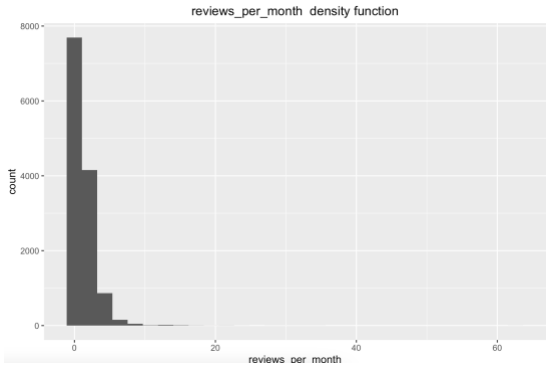
calculated_host_listings_count_private_home Boxplot



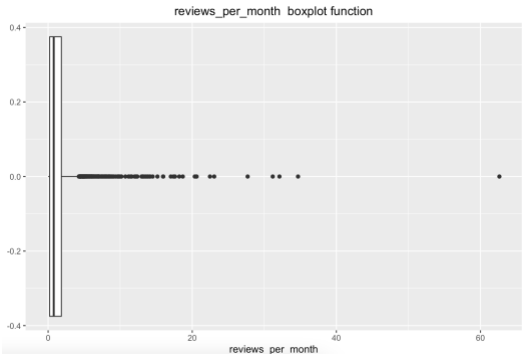
calculated_host_listings_count_shared_rooms Histogram



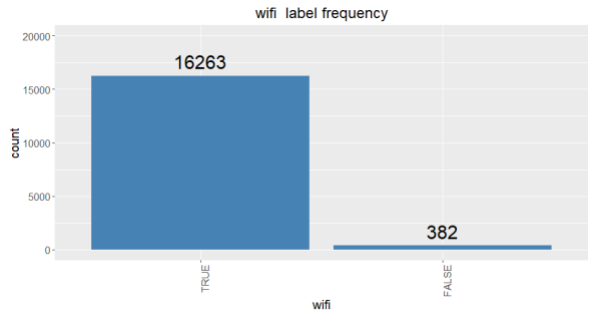
calculated_host_listings_count_shared_rooms Boxplot



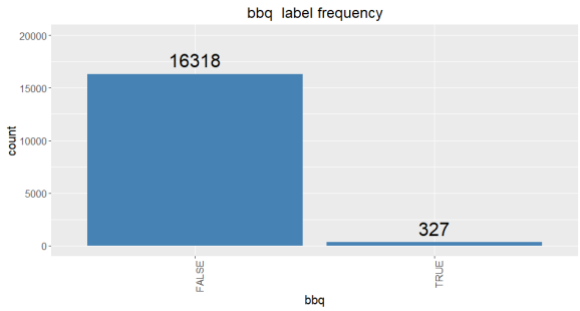
reviews_per_month Histogram



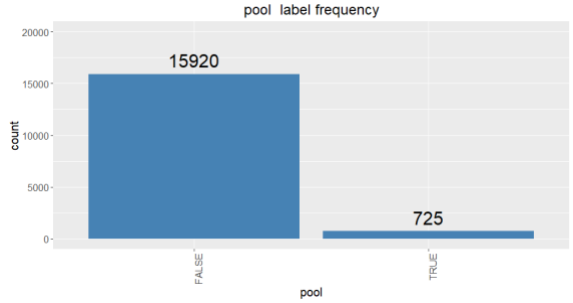
reviews_per_month Boxplot



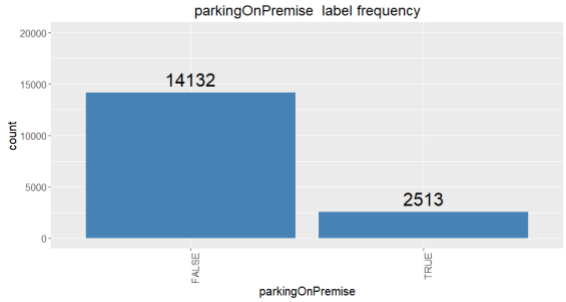
Barplot of wifi



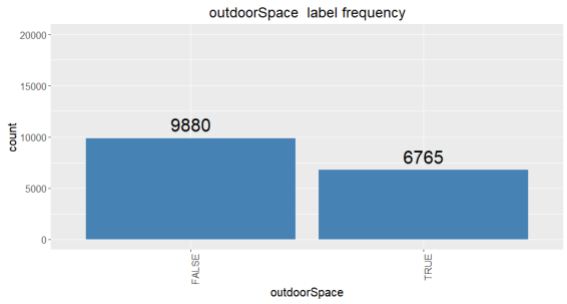
Barplot of bbq



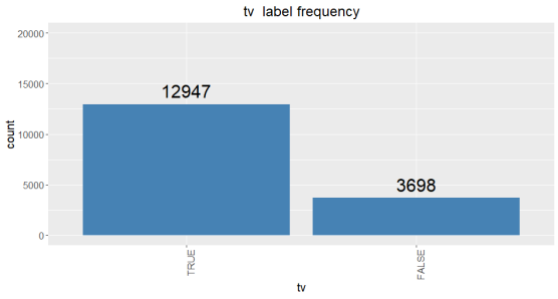
Barplot of pool



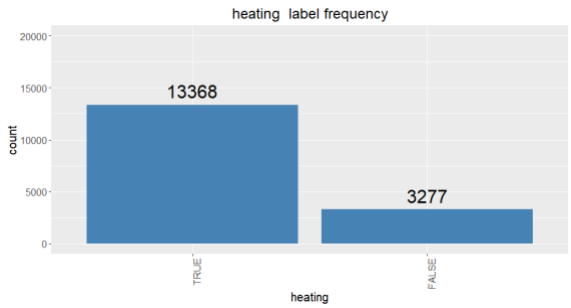
Barplot of parkingOnPremise



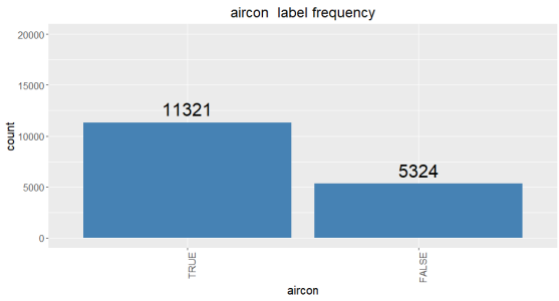
Barplot of outdoorSpace



Barplot of tv

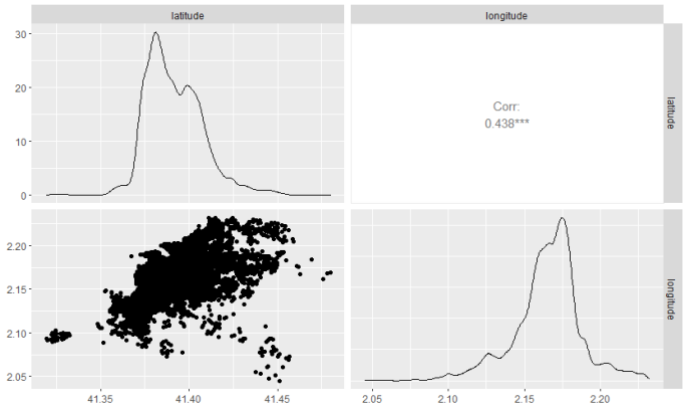


Barplot of heating

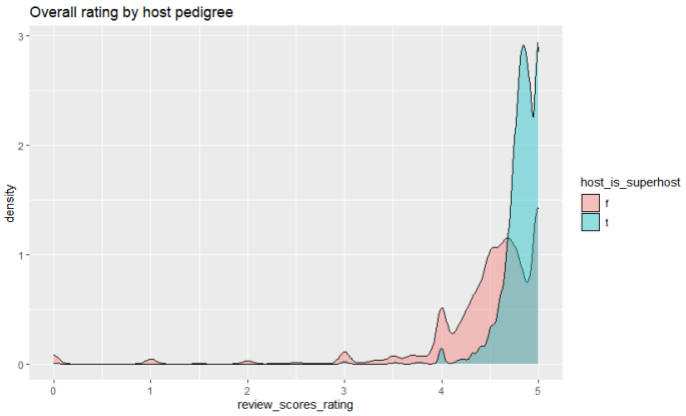


Barplot of aircon

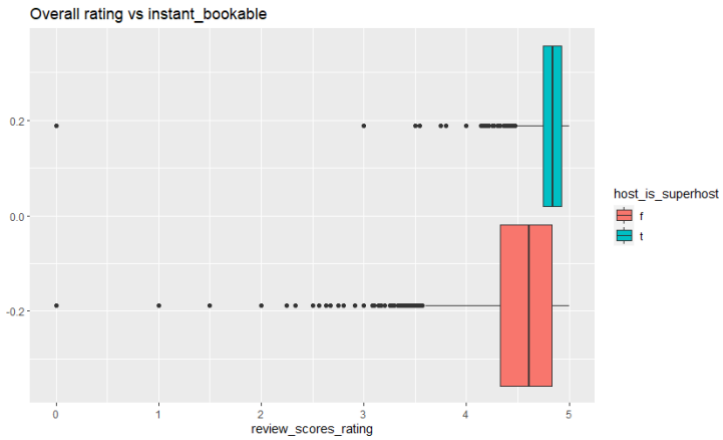
BIVARIATE ANALYSIS



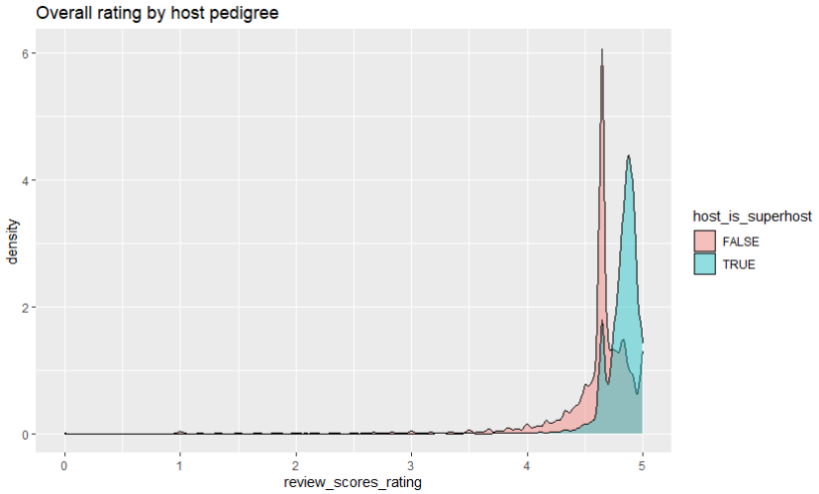
latitude and longitude correlation, biplot and density plots



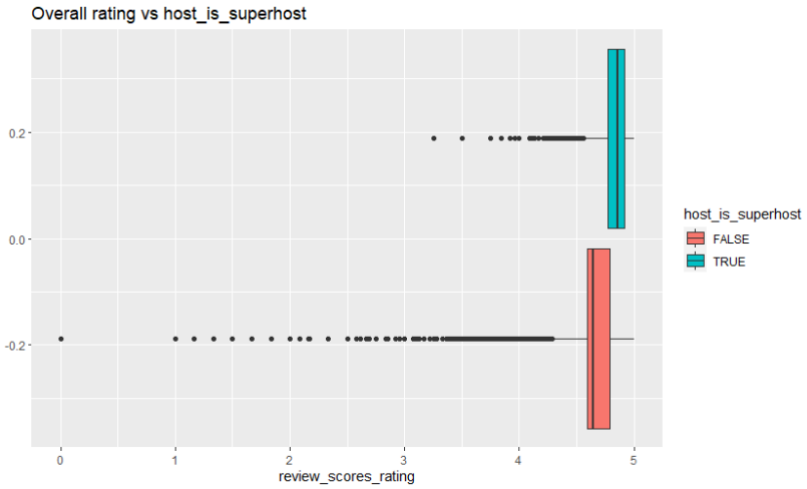
overall rating by host pedigree (old data)



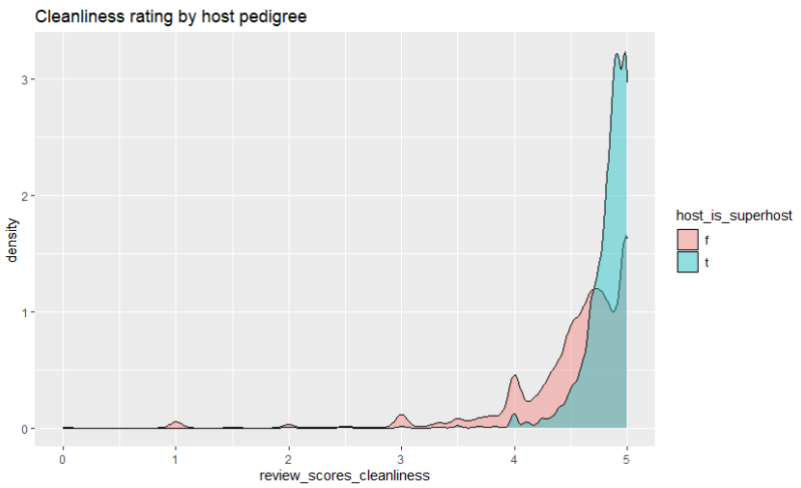
Boxplots comparing overall rating vs instant_bookable variables (old data)



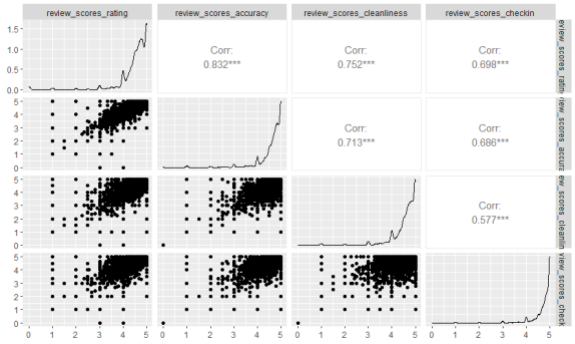
Density plots comparing overall rating vs host_is_superhost variables



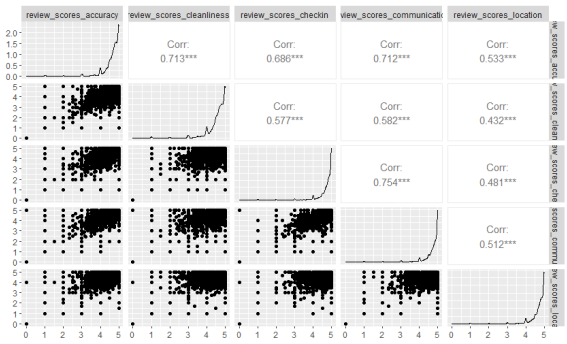
Boxplots comparing overall rating vs host_is_superhost variables



Density plots comparing review_scores_cleanliness vs host_is_superhost variables (old data)

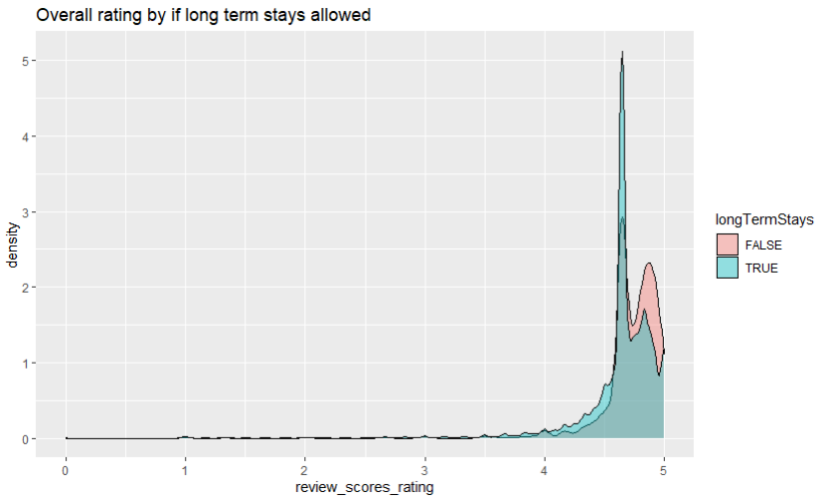


: Correlation between review score attributes

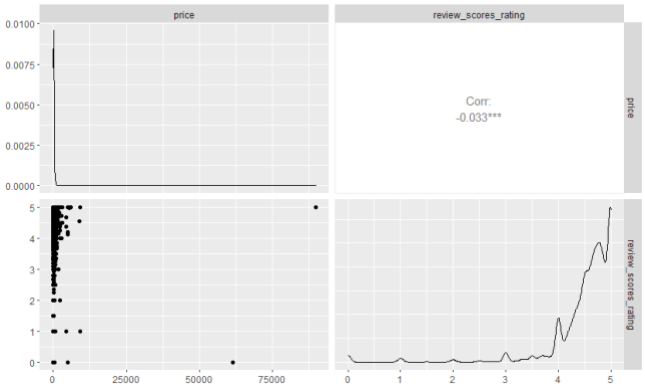


Correlation between review score attributes

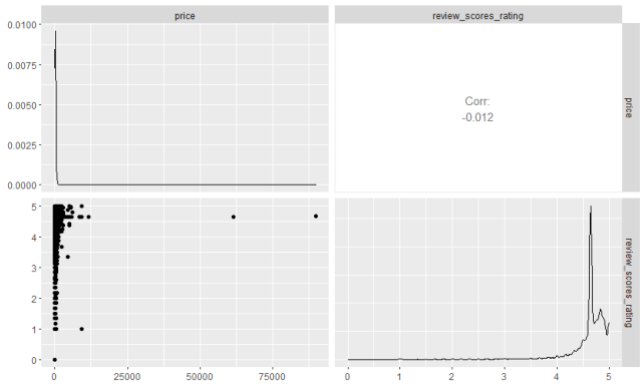
Price < 10000 vs host_is_superhost



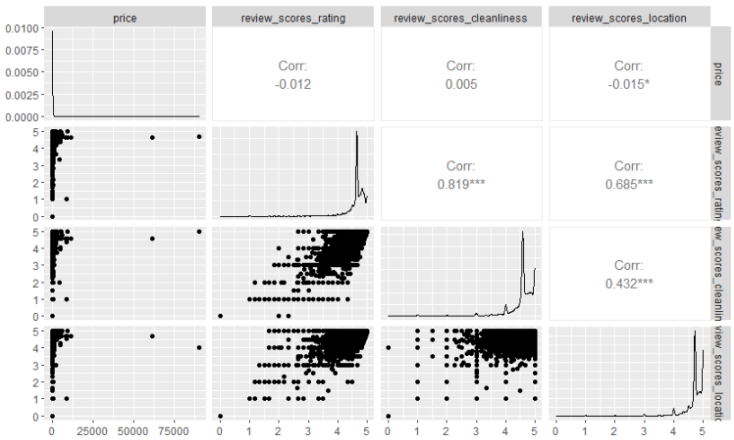
Overall rating vs longTermStays density



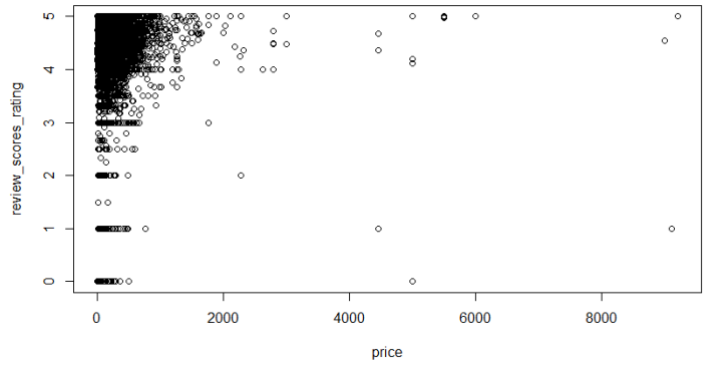
Correlation between price and review_scores_rating old dataset



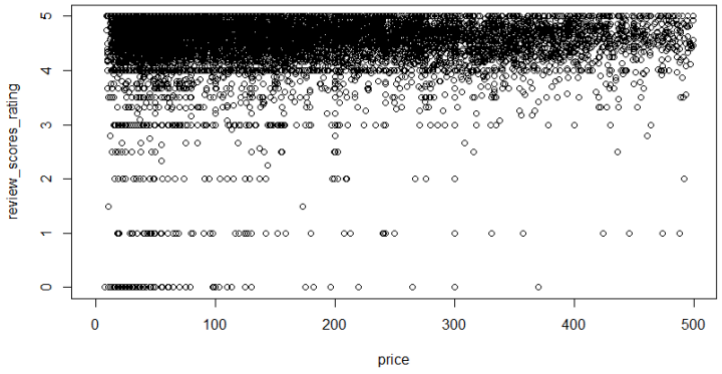
Correlation between price and review_scores_rating new dataset



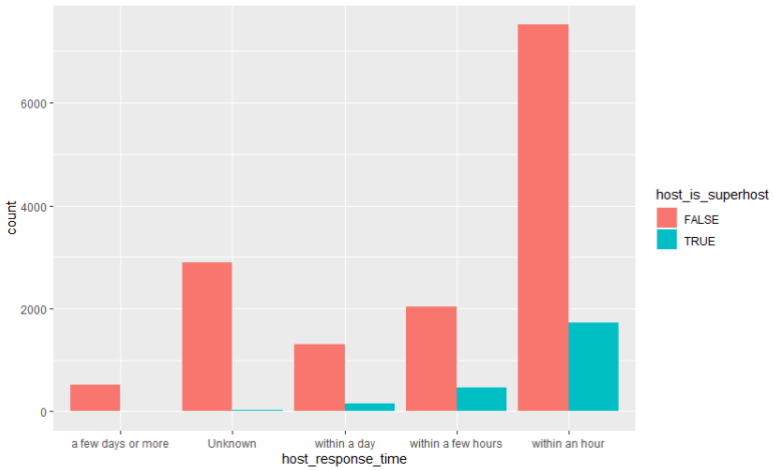
Correlation between price and review scores



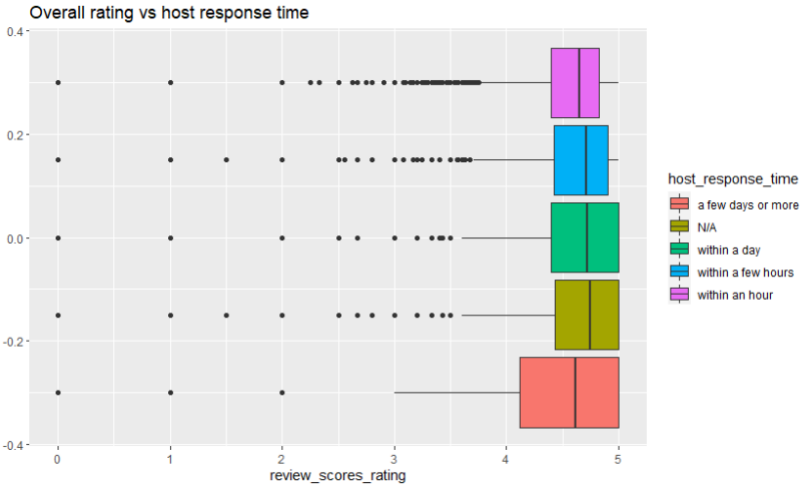
Scatterplot between price < 10000 and review_scores_rating



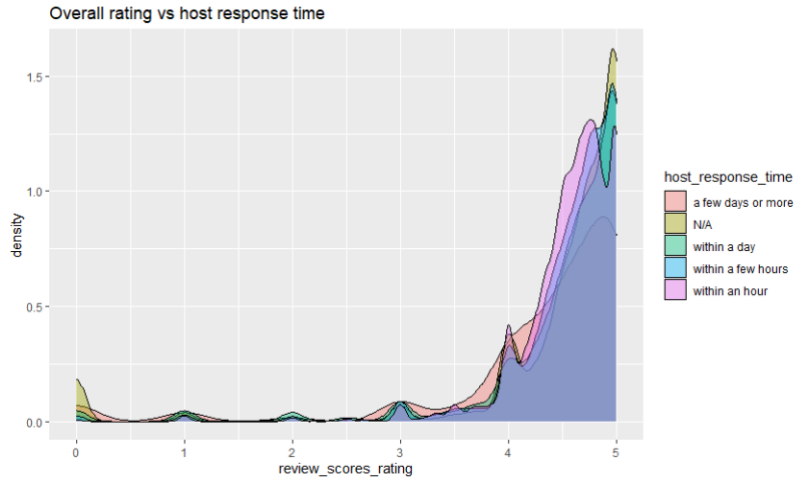
Scatterplot between price < 500 and review_scores_rating



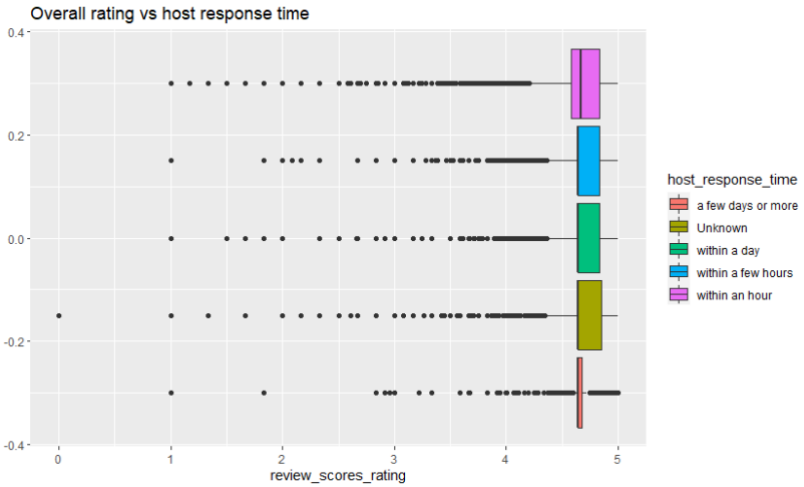
Host response time and host is superhost biplot



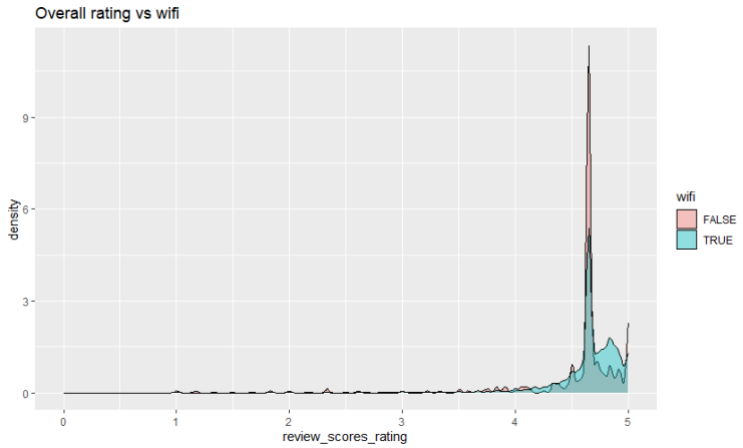
Review scores rating and host response time boxplot



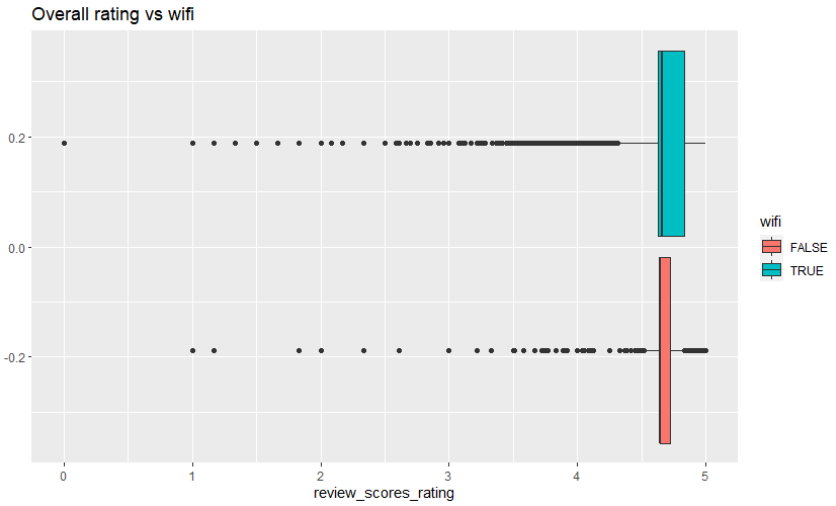
Review scores rating and host response time density plot (old dataset)



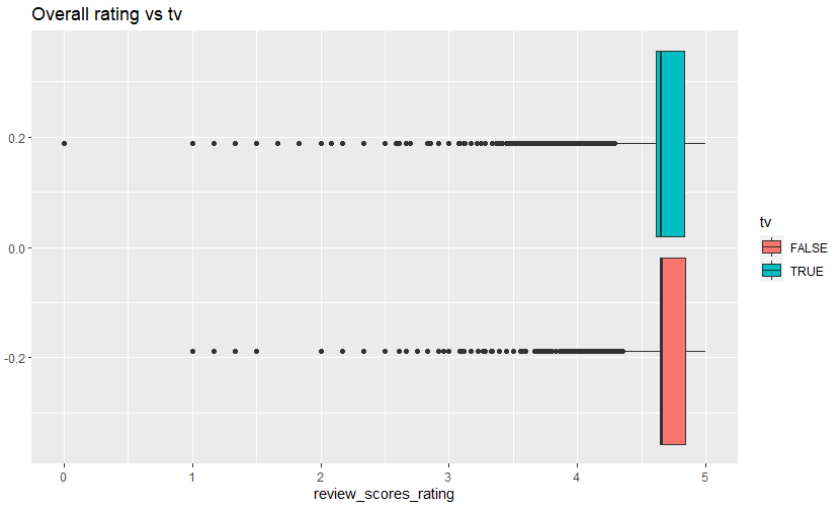
Review scores rating and host response time boxplot (old dataset)



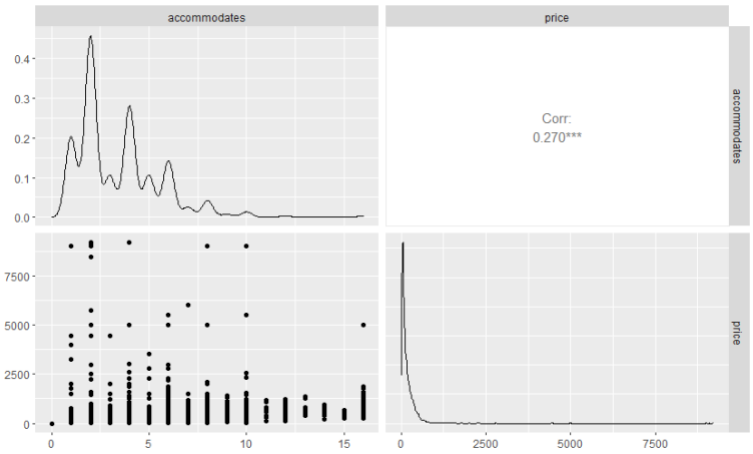
Review scores rating and wifi plot



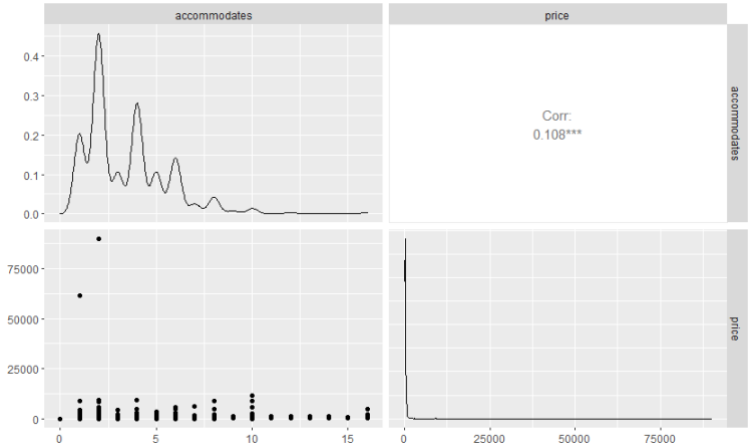
Review scores rating and wifi boxplot



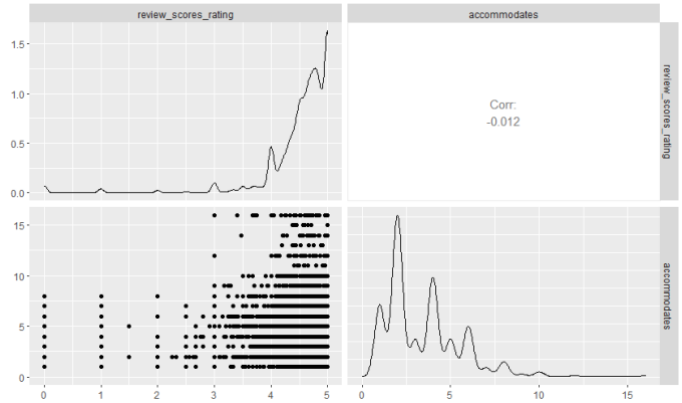
Review scores rating and tv boxplot



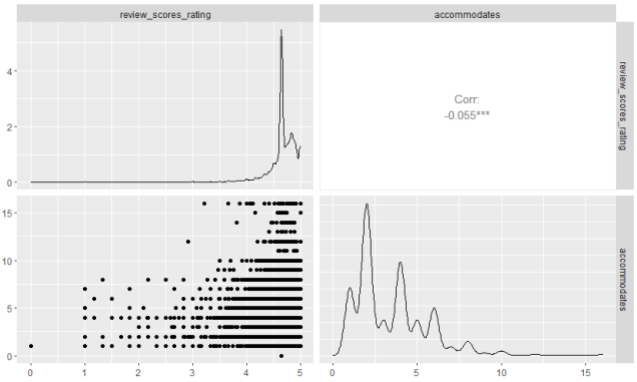
Accommodates and price correlation, biplot and density plots (price is from old dataset)



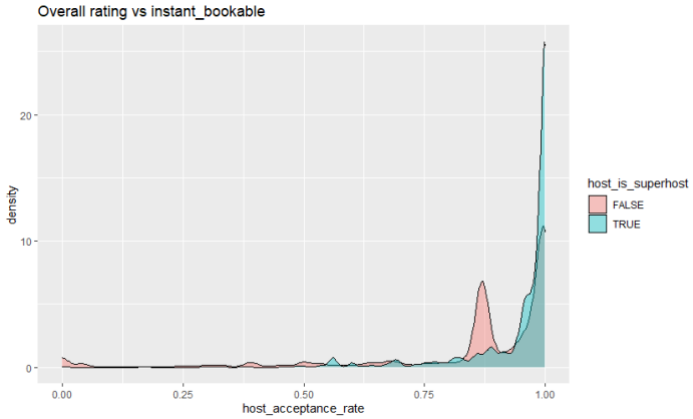
Accommodates and price correlation, biplot and density plots (price is from old dataset)



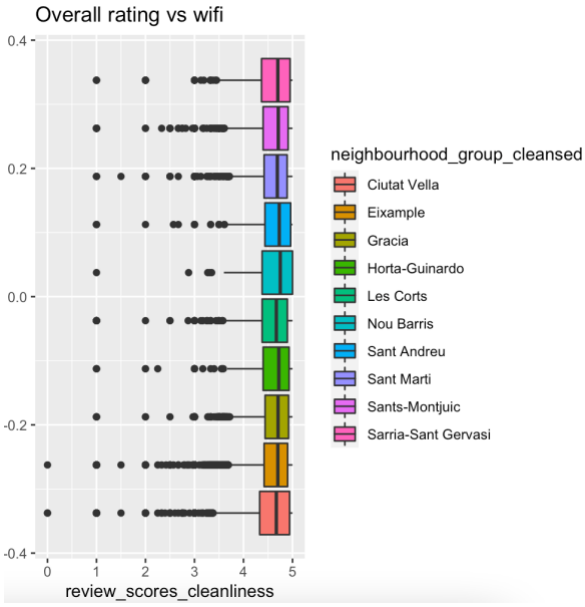
Review scores rating and accommodates correlation, biplot and density plots



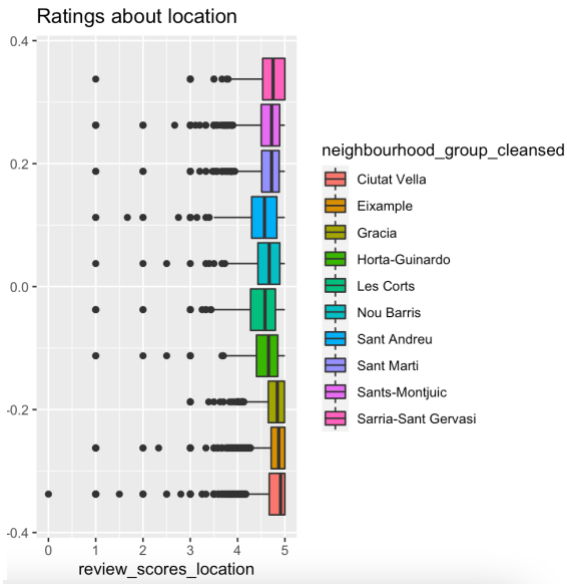
Review_scores_rating and accommodates correlation, biplot and density plots



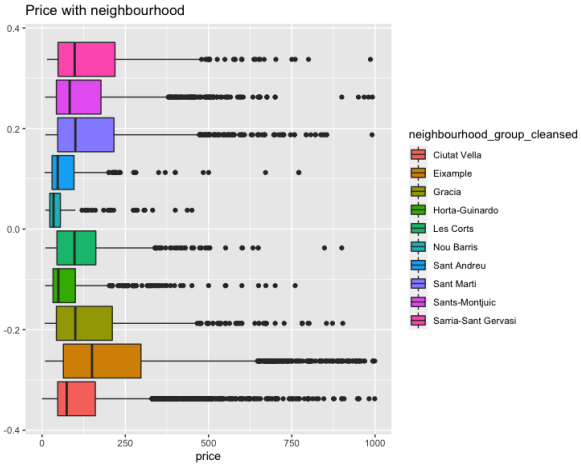
Host acceptance rate and host is superhost density plot



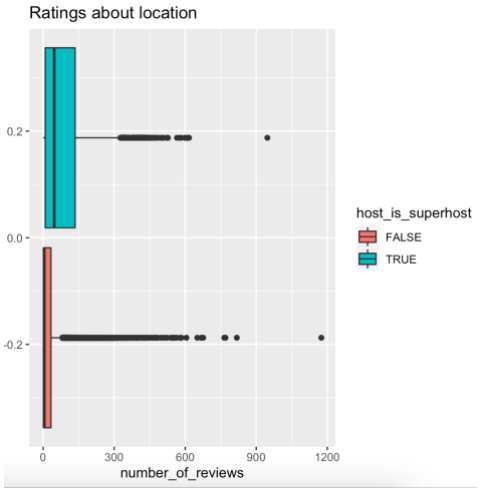
Cleanliness score about location



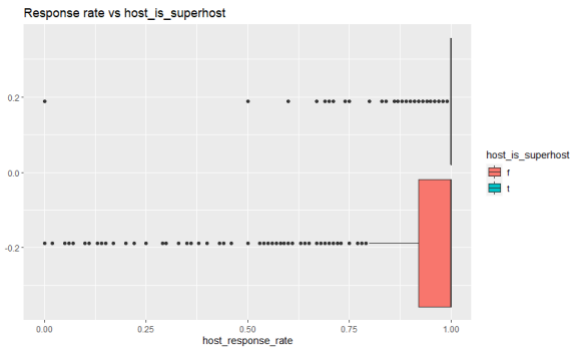
Location score about location



Number of reviews by host pedigree

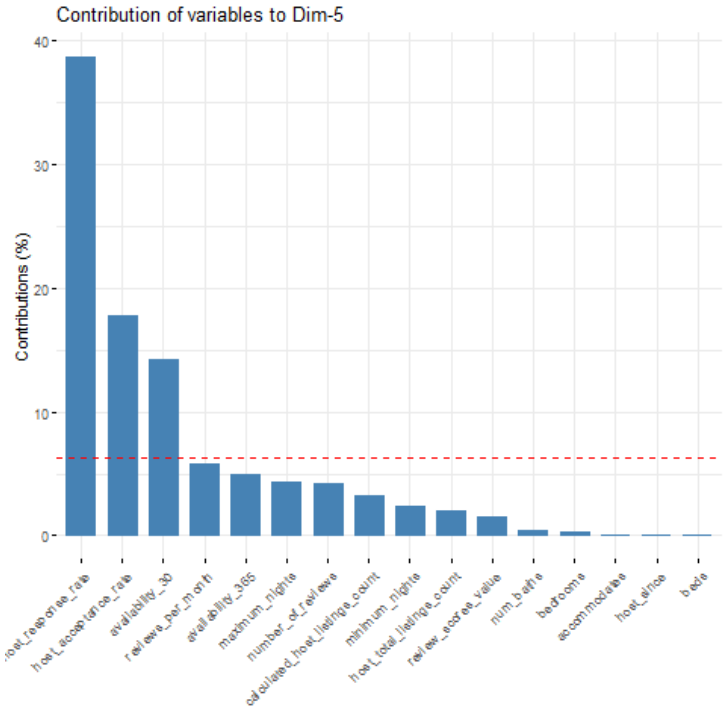


Number of reviews by host pedigree

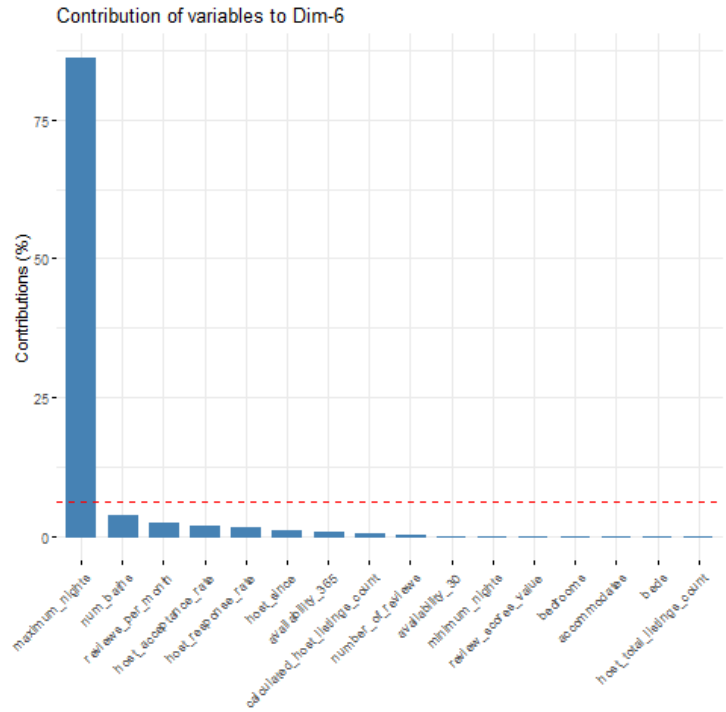


Host response rate by host pedigree

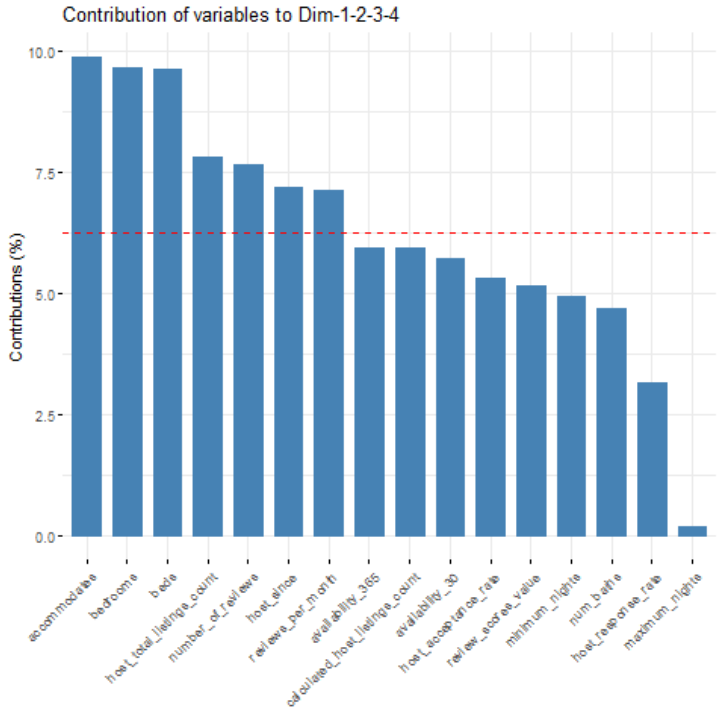
PCA



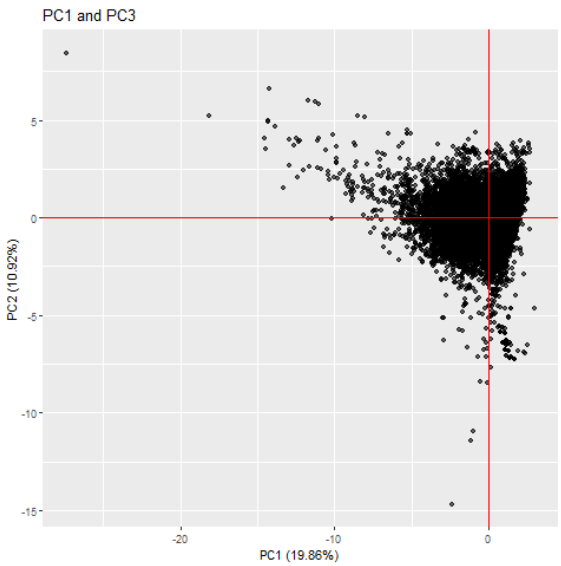
Contribution of variables to PC5



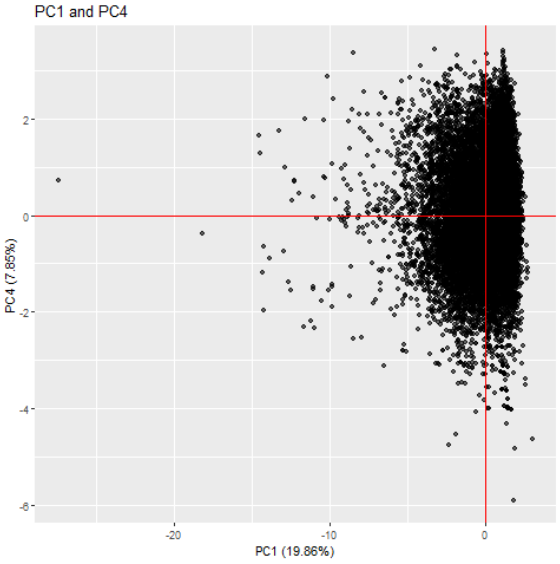
Contribution of variables to PC6



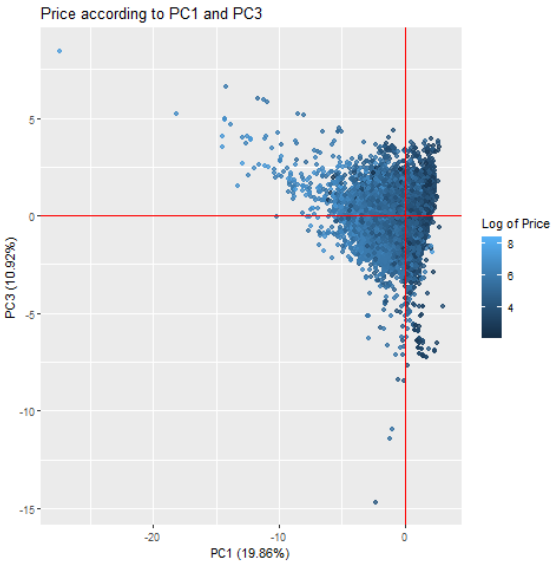
Contribution of variables to PC1, PC2, PC3 and PC4.



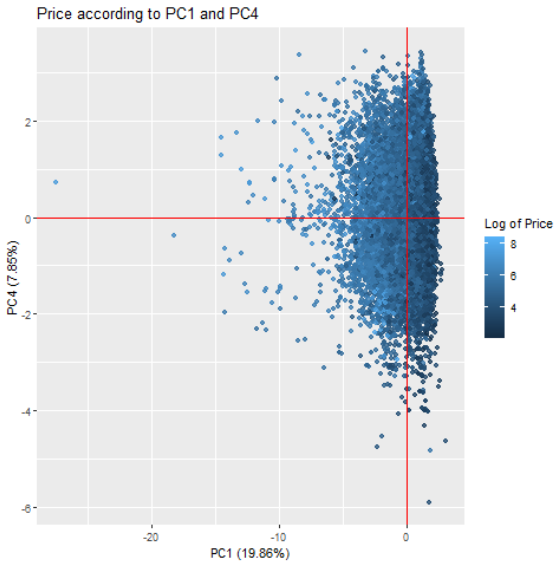
Individuals factor map according to PC1 and PC4



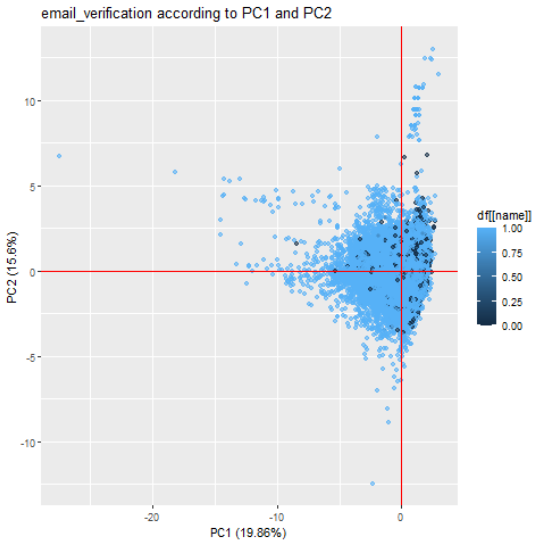
Individuals factor map according to PC1 and PC4



Individuals factor map according to PC1 and PC3 with price as a supplementary variable.

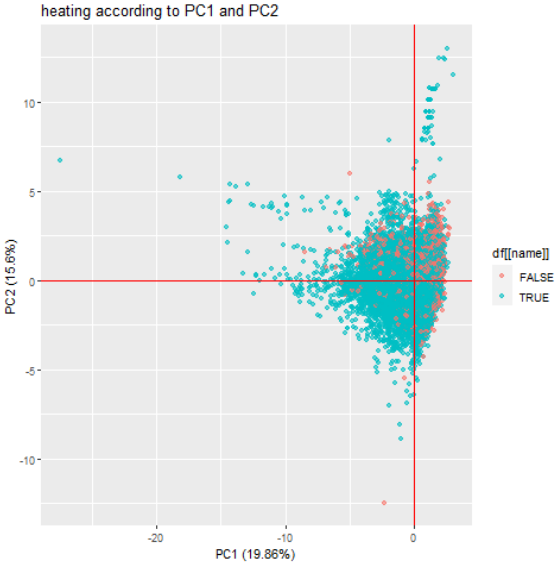


Individuals factor map according to PC1 and PC4 With price as a supplementary variable.



Individuals factor map according to PC1 and PC2 with email_verification as a supplementary variable.

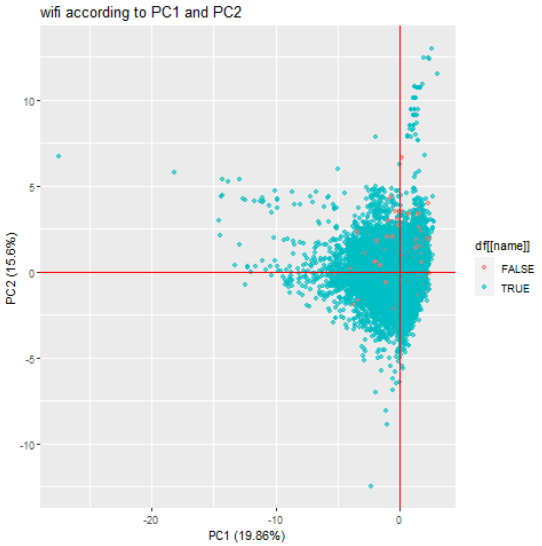
Individuals factor map according to PC1 and PC2 with hairDryer as a supplementary variable.



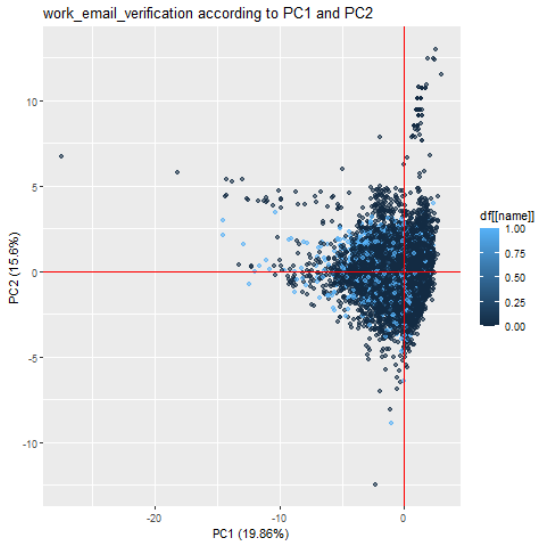
Individuals factor map according to PC1 and PC2
with heating as a supplementary variable.

Individuals factor map according to PC1 and PC2
with host_has_profile_pic as a supplementary variable.

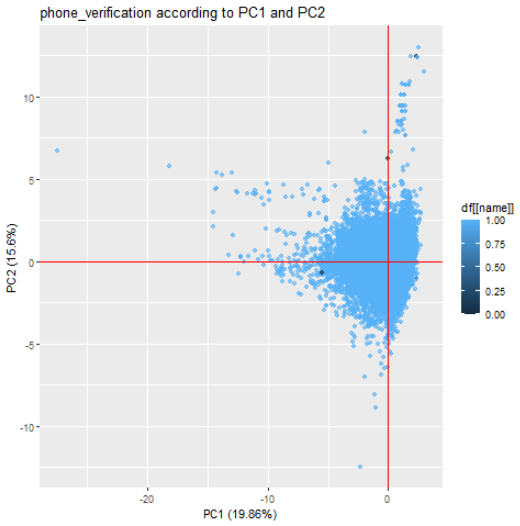
Individuals factor map according to PC1 and PC2
With host_identity_verified as a supplementary variable.



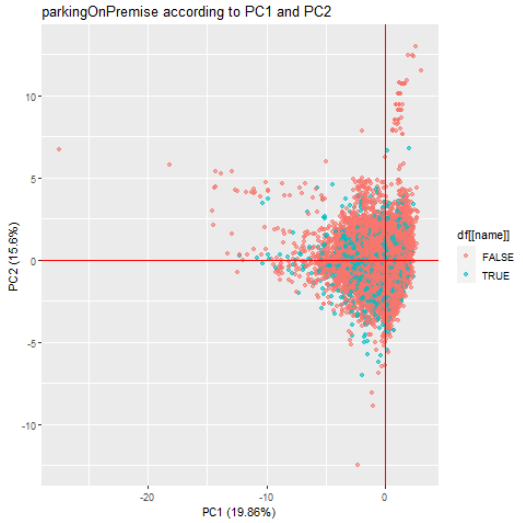
Individuals factor map according to PC1 and PC2
with wifi as a supplementary variable.



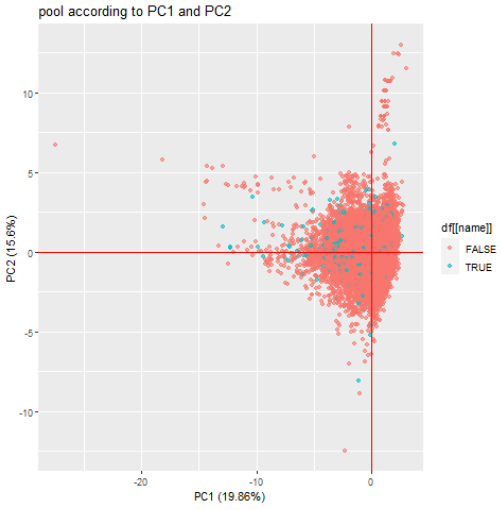
Individuals factor map according to PC1 and PC2 with work_email_verification as a supplementary variable.



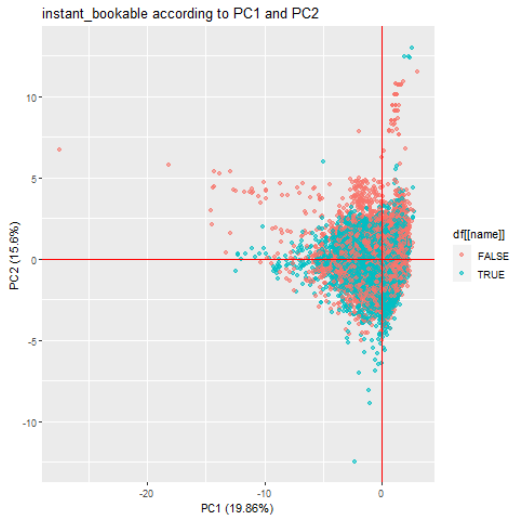
I Individuals factor map according to PC1 and PC2 with phone_verification as a supplementary variable.



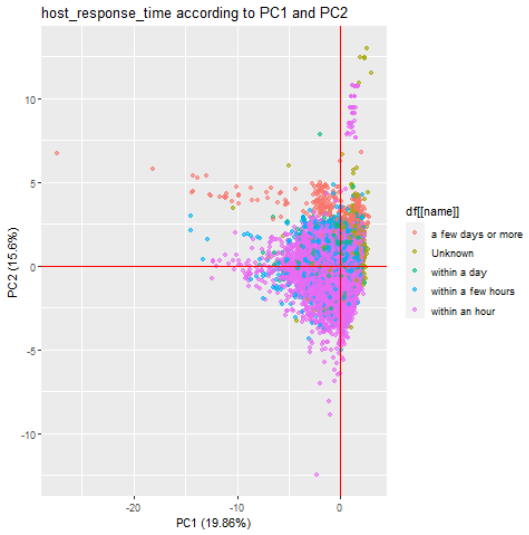
Individuals factor map according to PC1 and PC2 with parkingOnPremises as a supplementary variable.



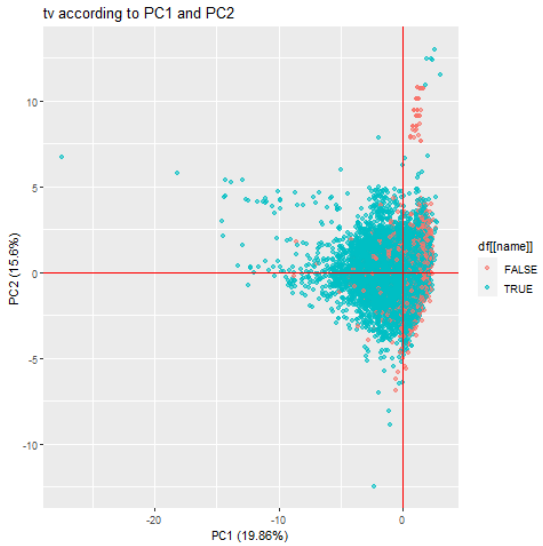
Individuals factor map according to PC1 and PC2 with pool as a supplementary variable.



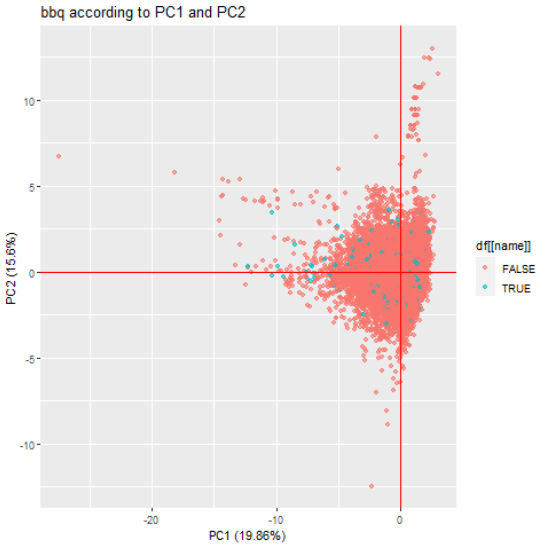
Individuals factor map according to PC1 and PC2 with instant_bookable as a supplementary variable.



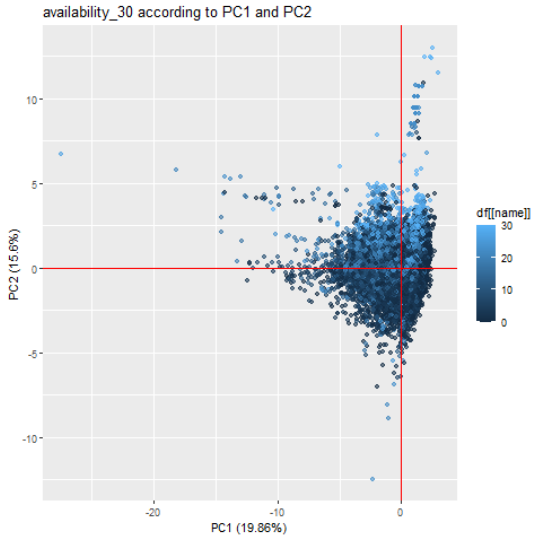
Individuals factor map according to PC1 and PC2 with host_response_time as a supplementary variable.



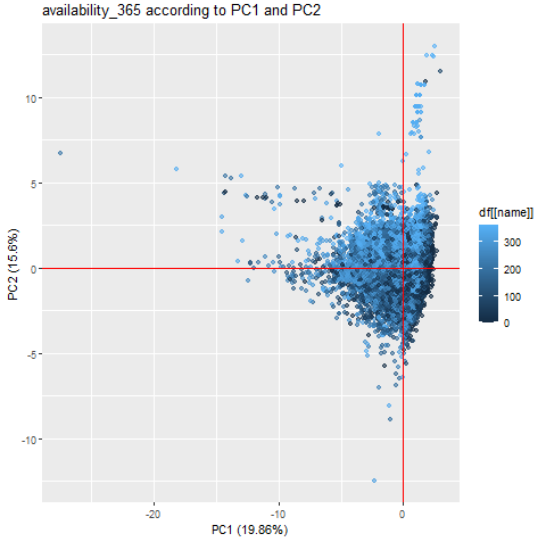
Individuals factor map according to PC1 and PC2 with tv as a supplementary variable.



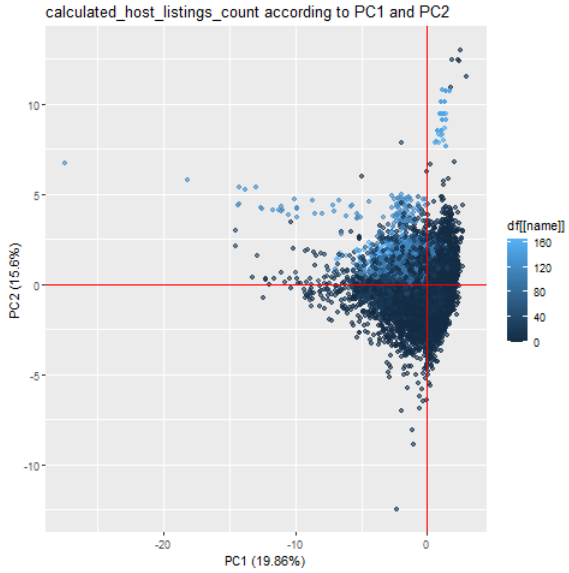
Individuals factor map according to PC1 and PC2 with bbq as a supplementary variable.



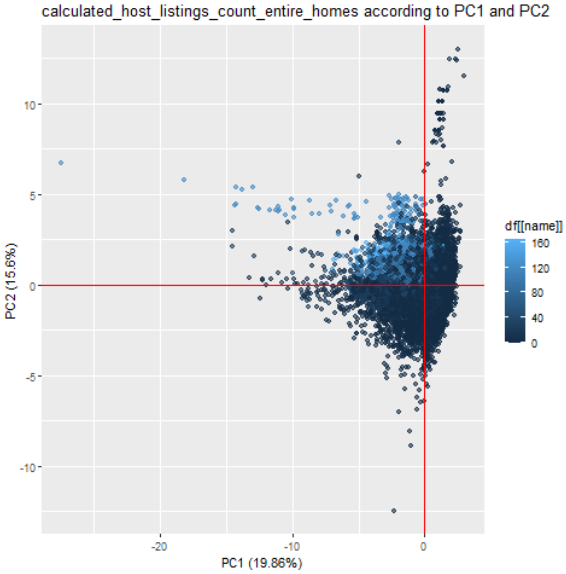
Individuals factor map according to PC1 and PC2 with availability_30 as a supplementary variable.



Individuals factor map according to PC1 and PC2 with availability_365 as a supplementary variable.



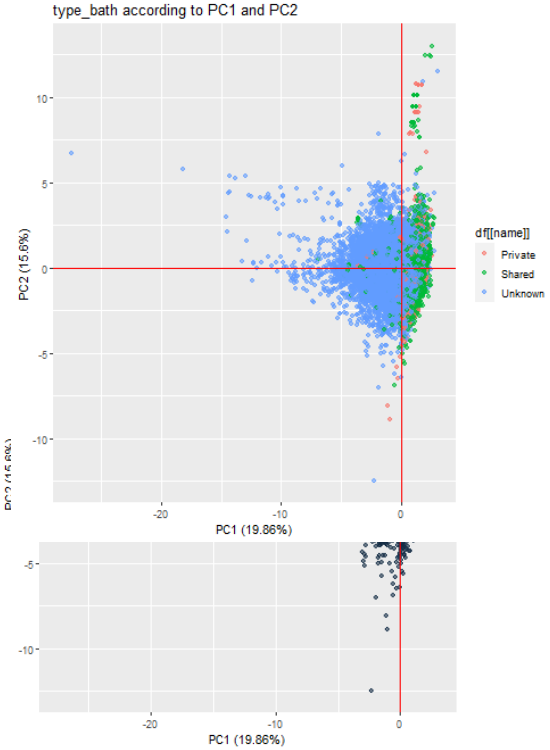
Individuals factor map according to PC1 and PC2 with calculated_host_listings_count as a supplementary variable.



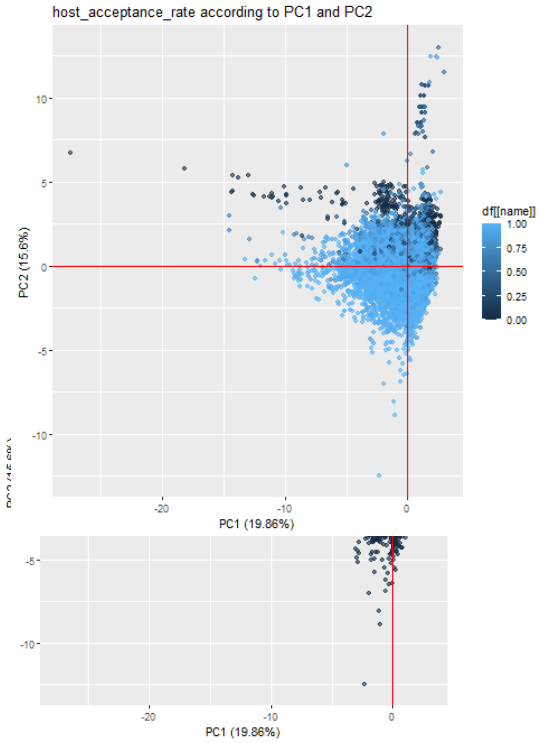
Individuals factor map according to PC1 and PC2 with calculated_host_listings_count_entire_homes as supplementary.

Individuals factor map according to PC1 and PC2 with calculated_host_listings_count_private_rooms as a supplementary variable.

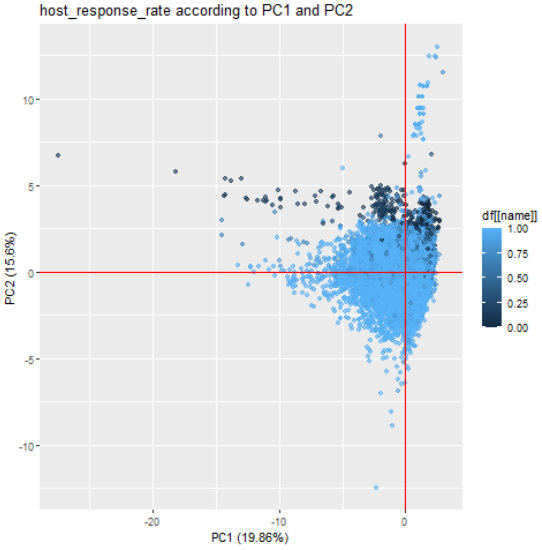
Individuals factor map according to PC1 and PC2 with calculated_host_listings_count_shared_rooms as a supplementary variable.



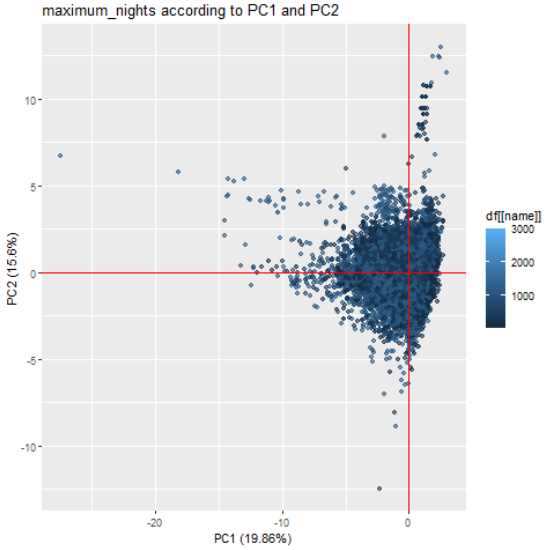
Individuals factor map according to PC1 and PC2 with type_bath as a supplementary variable.



Individuals factor map according to PC1 and PC2 with host_acceptance_rate as a supplementary variable.

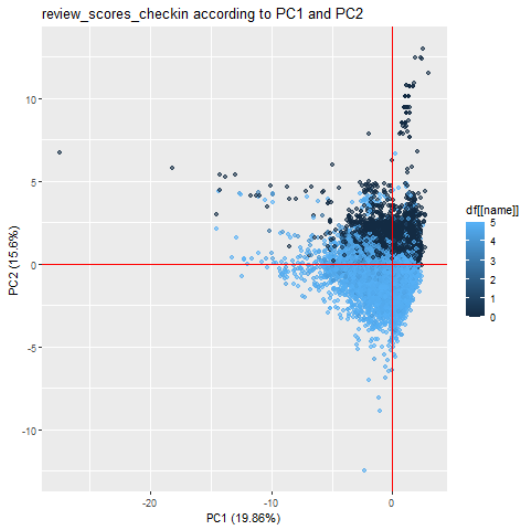


Individuals factor map according to PC1 and PC2
with host_response_rate as a supplementary variable.



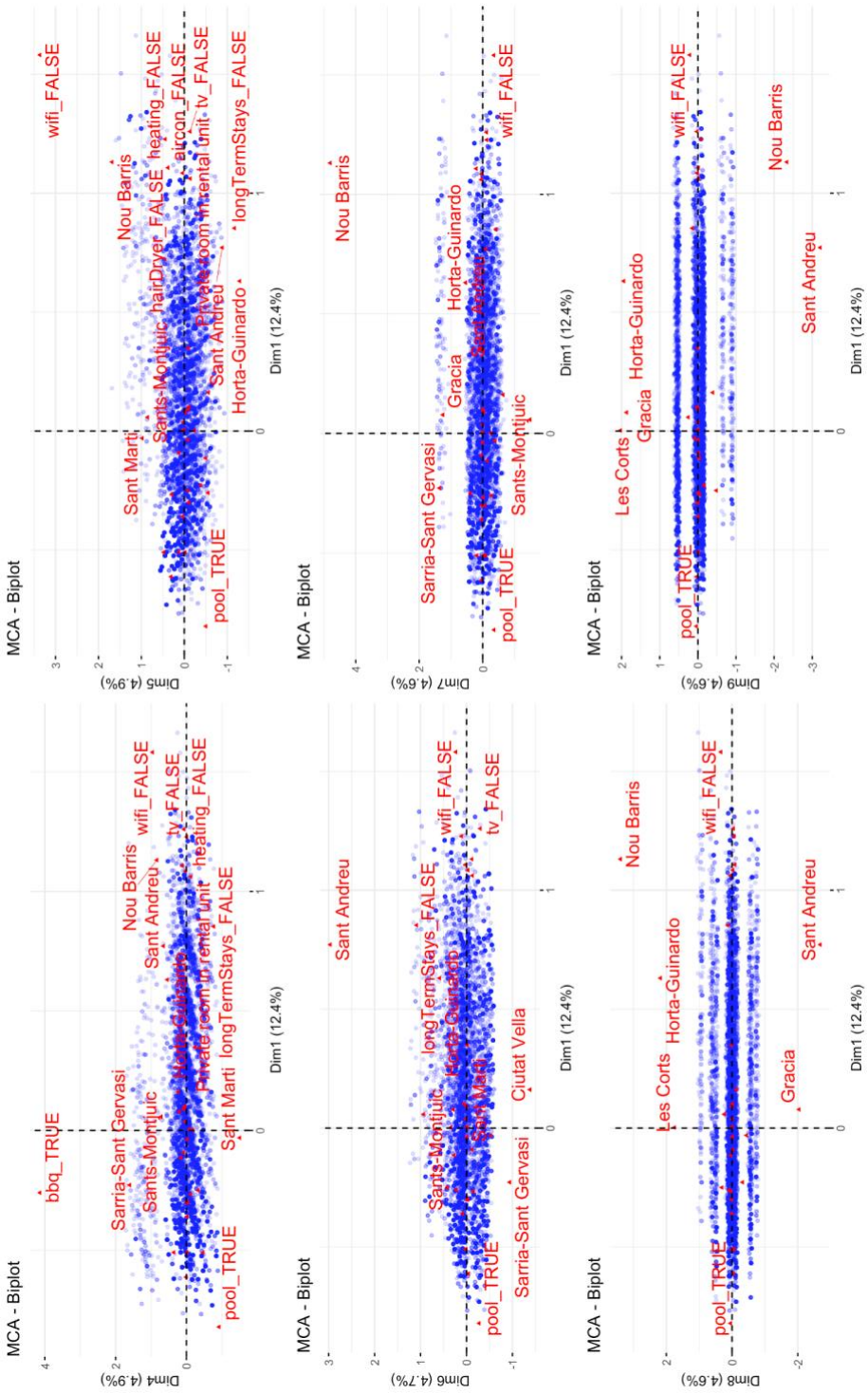
Individuals factor map according to PC1 and PC2
with maximum_nights as a supplementary variable.

Individuals factor map according to PC1 and PC2
with review_scores_accuracy as a supplementary variable.

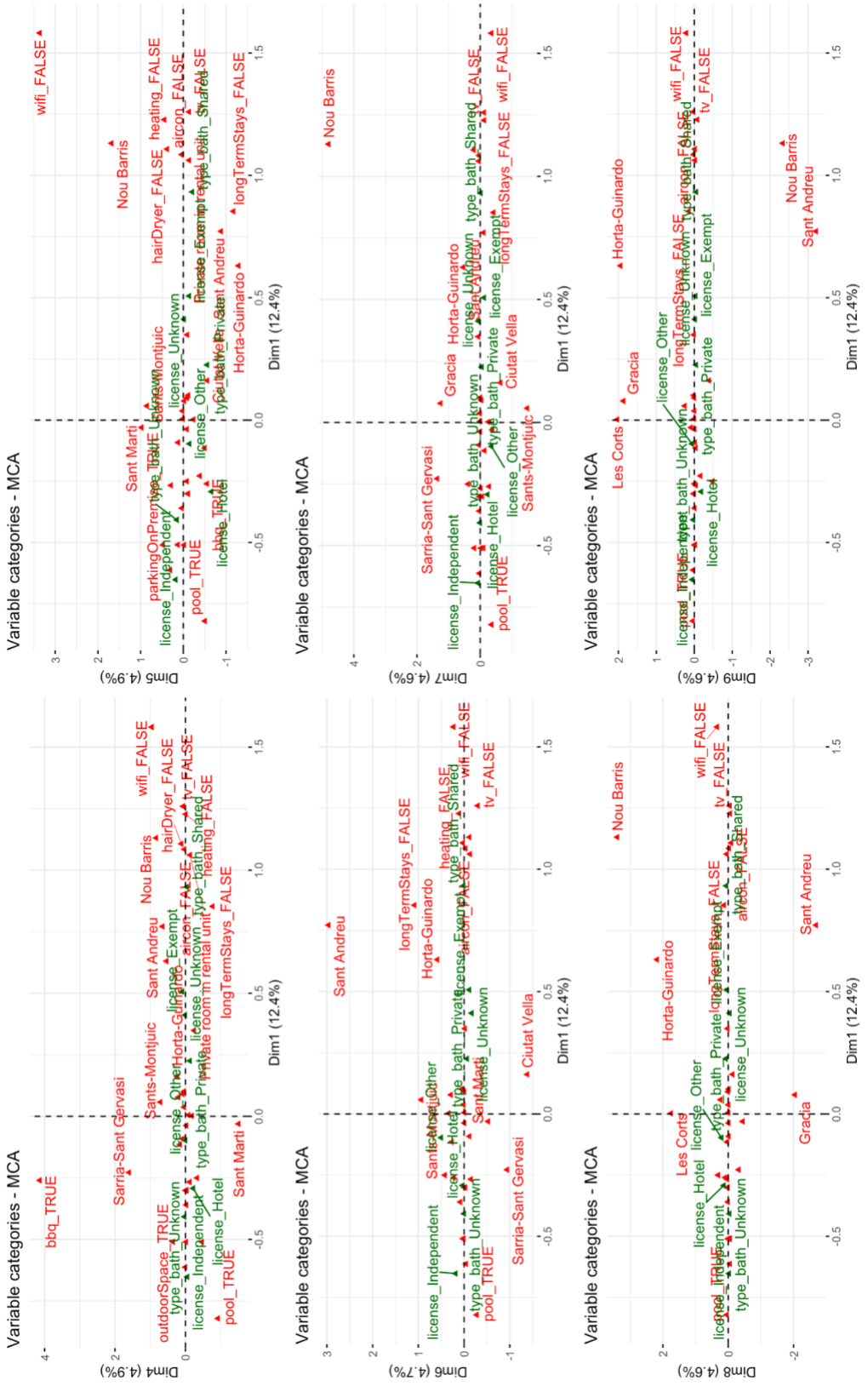


Individuals factor map according to PC1 and PC2
with review_scores_checkin as a supplementary variable.

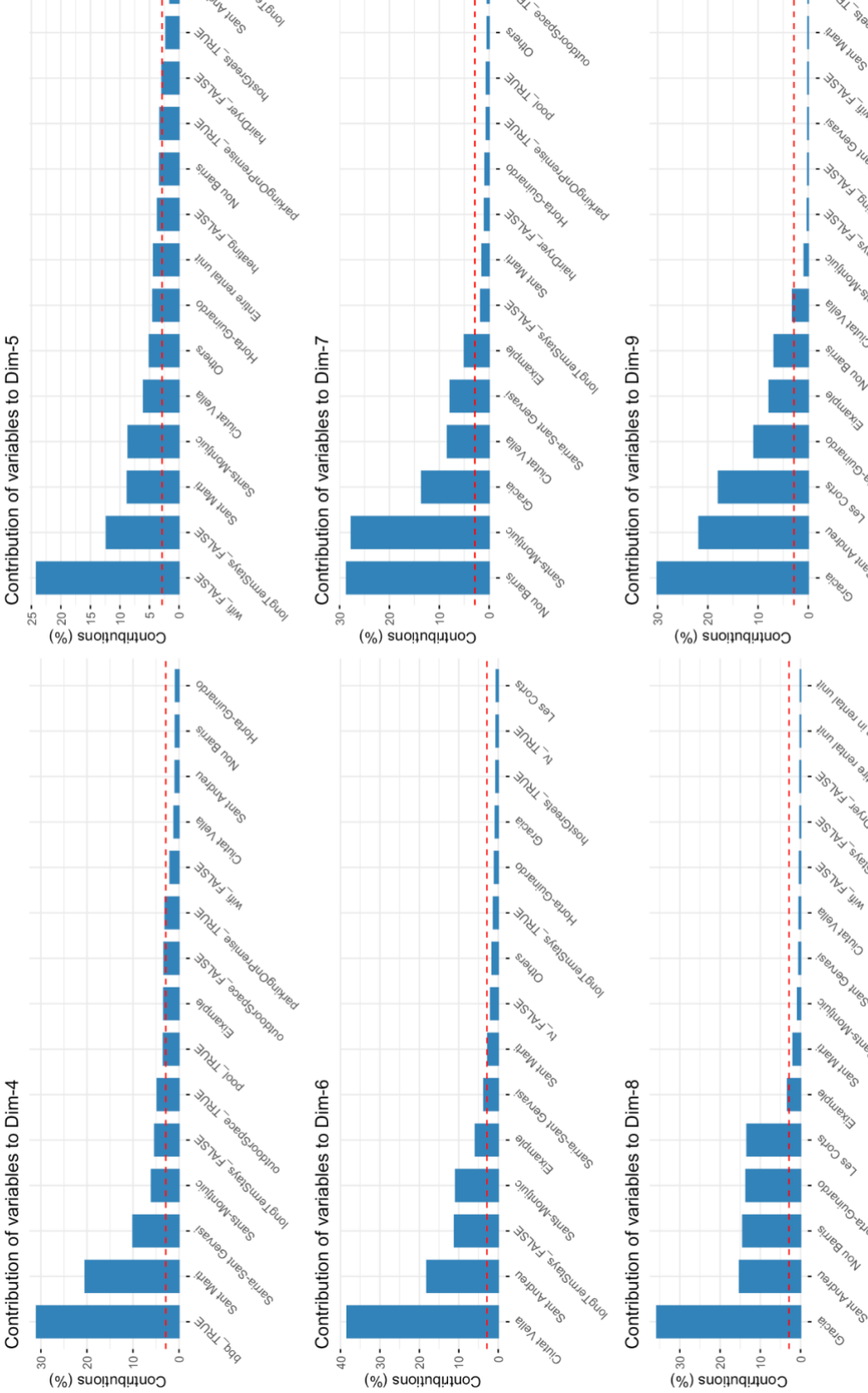
MCA



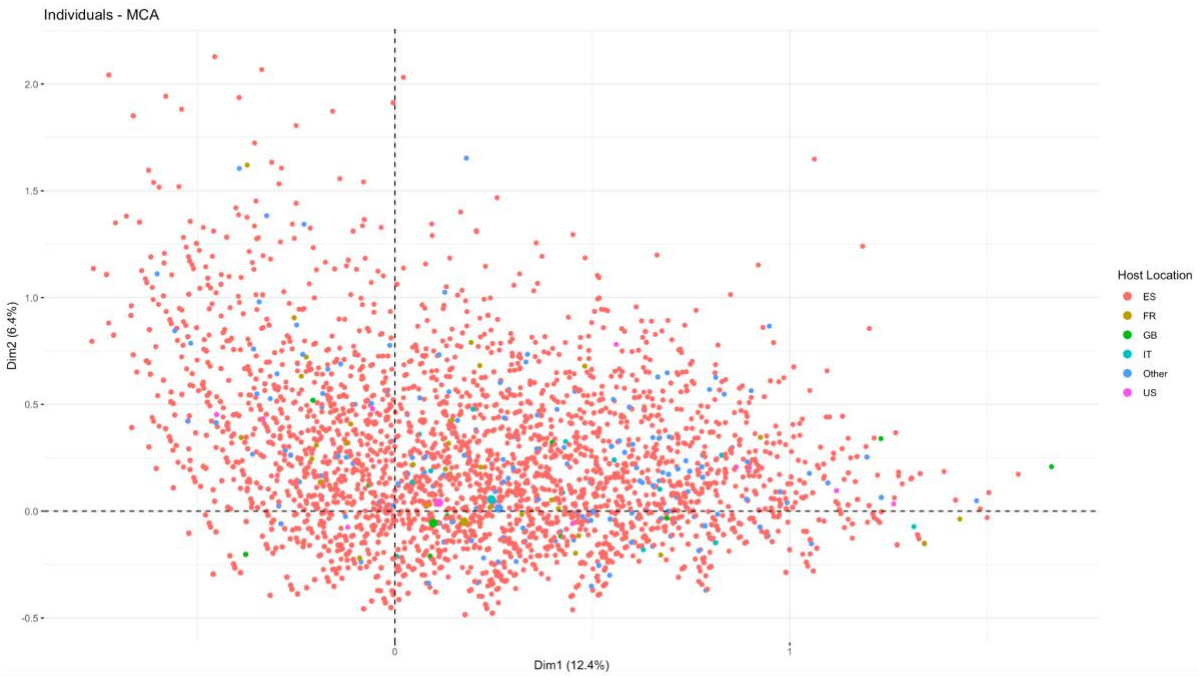
Biplot variables / individuals in dimensions 1 and 4 to 9



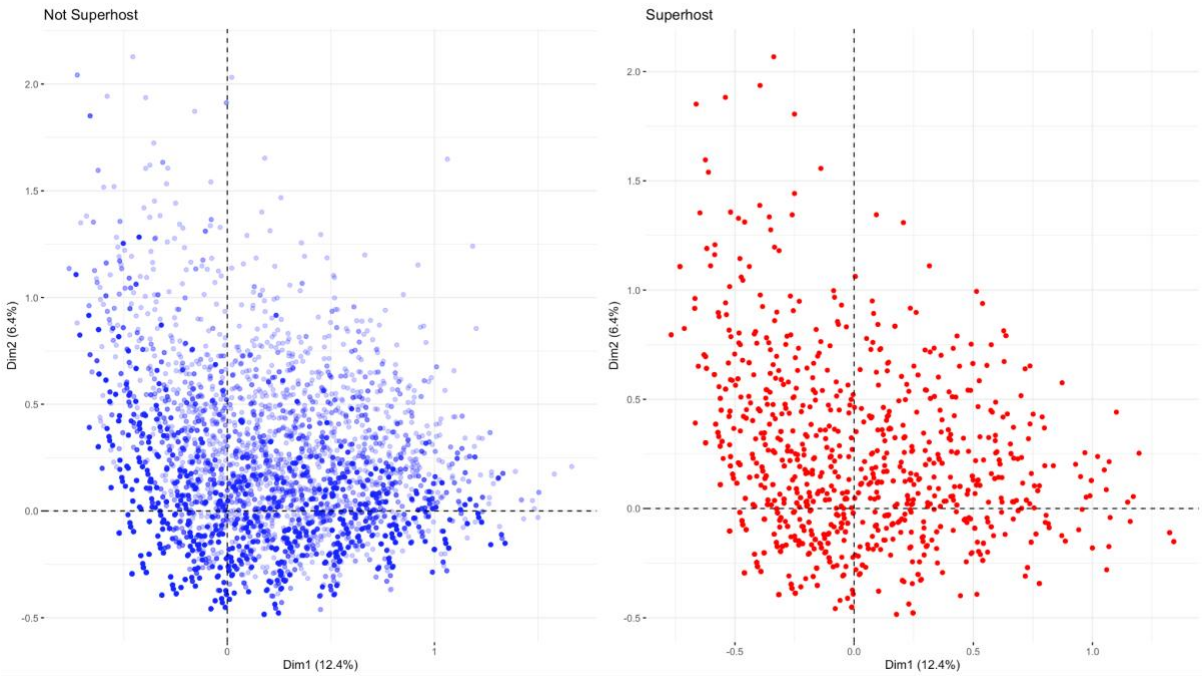
Correlation between variables in dimensions 1 and 4 to 9 with supplementary variables



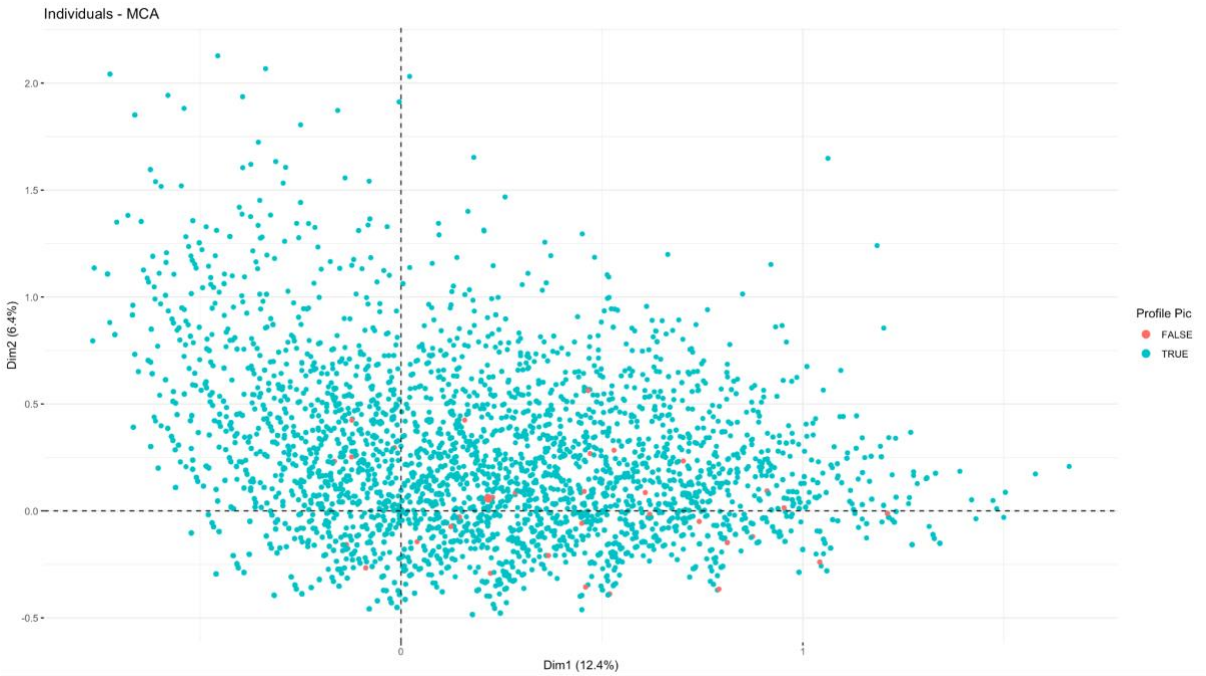
Contributions in dimensions from 4 to 9



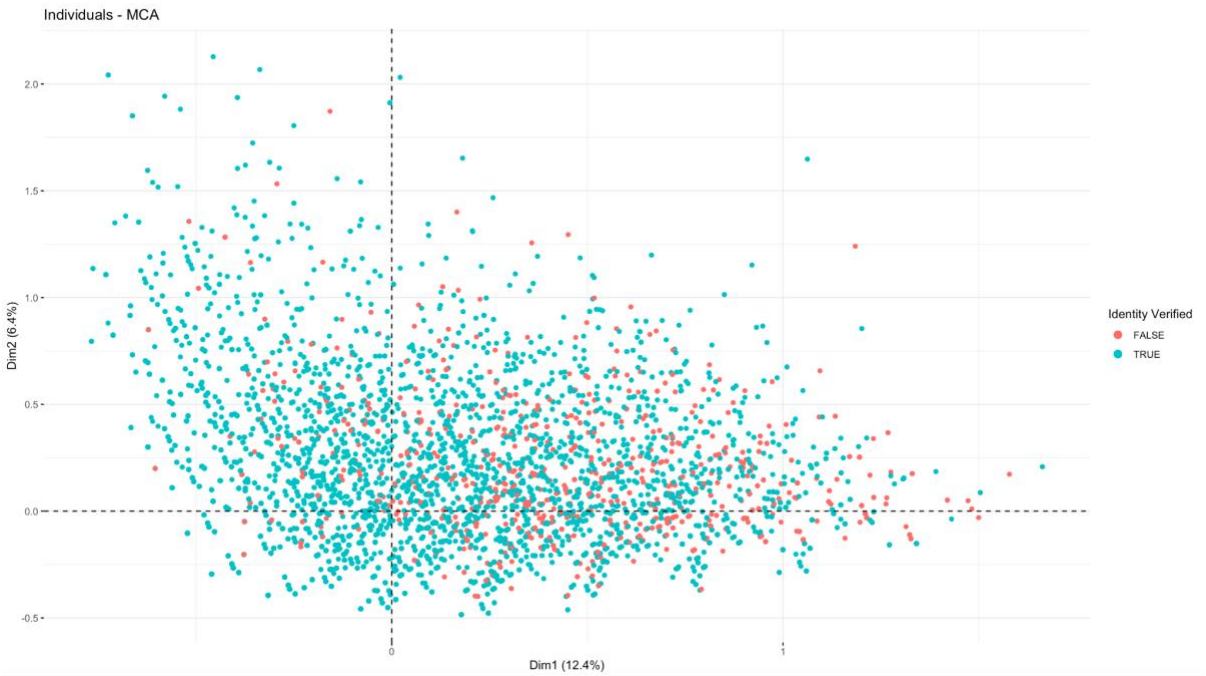
Host location dimensions 1 and 2



Host is superhost dimensions 1 and 2

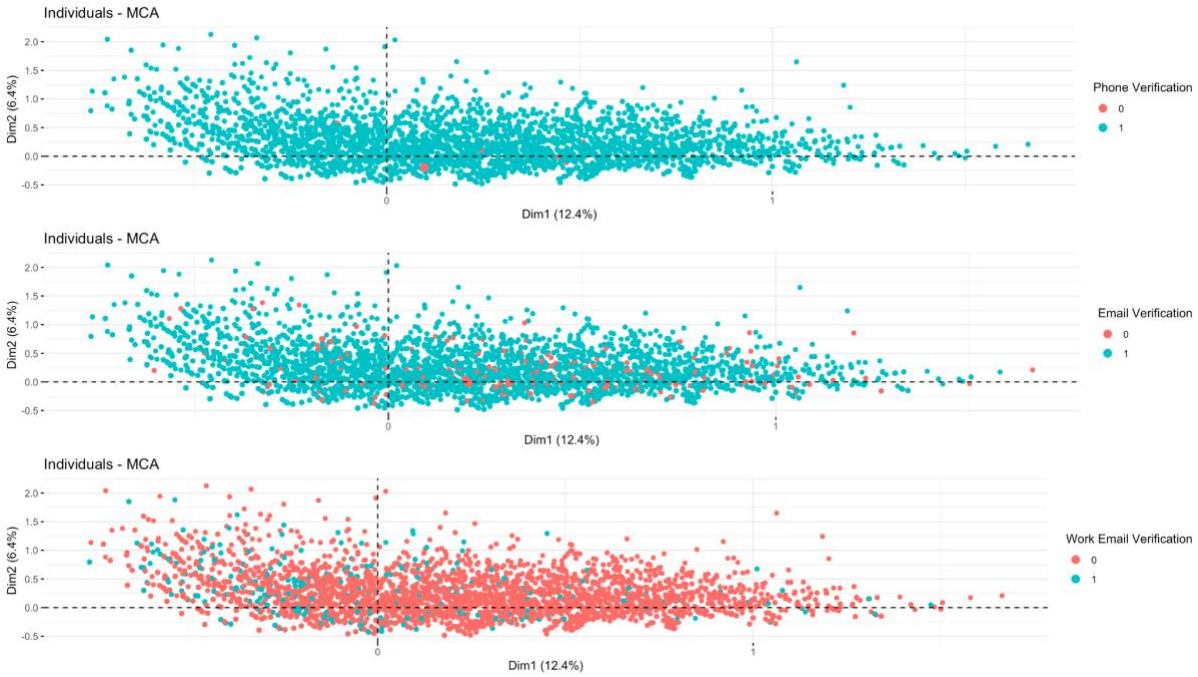


Host has profile picture dimensions 1 and 2

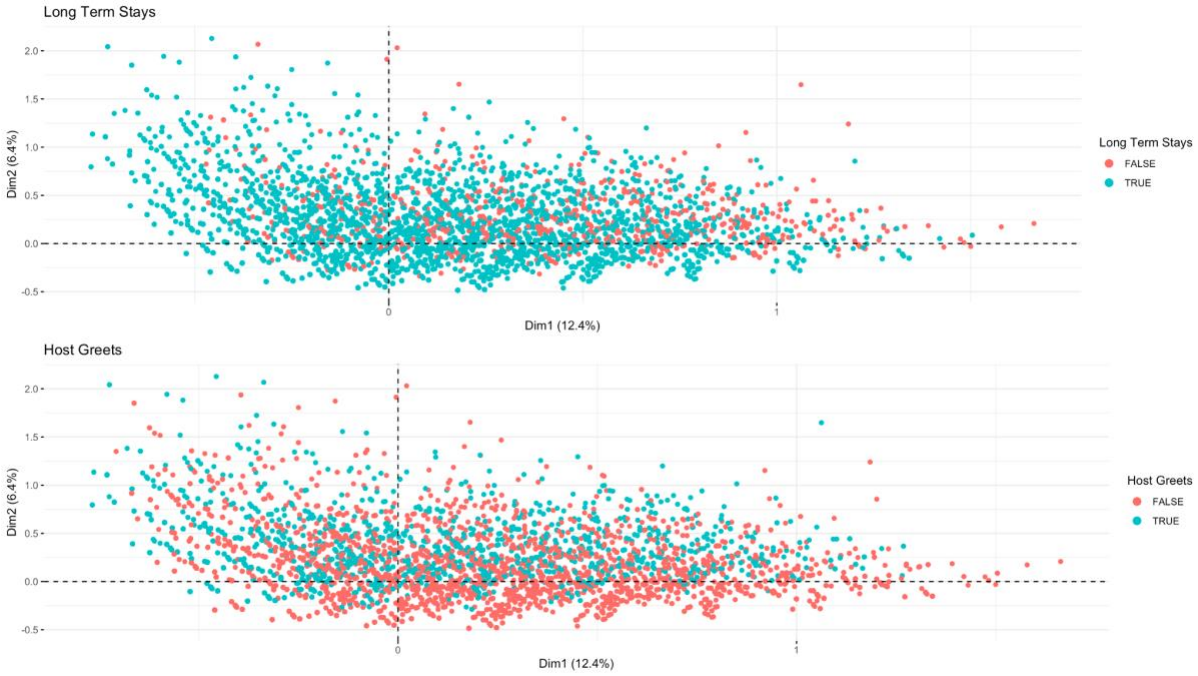


Host identity verified dimensions 1 and 2

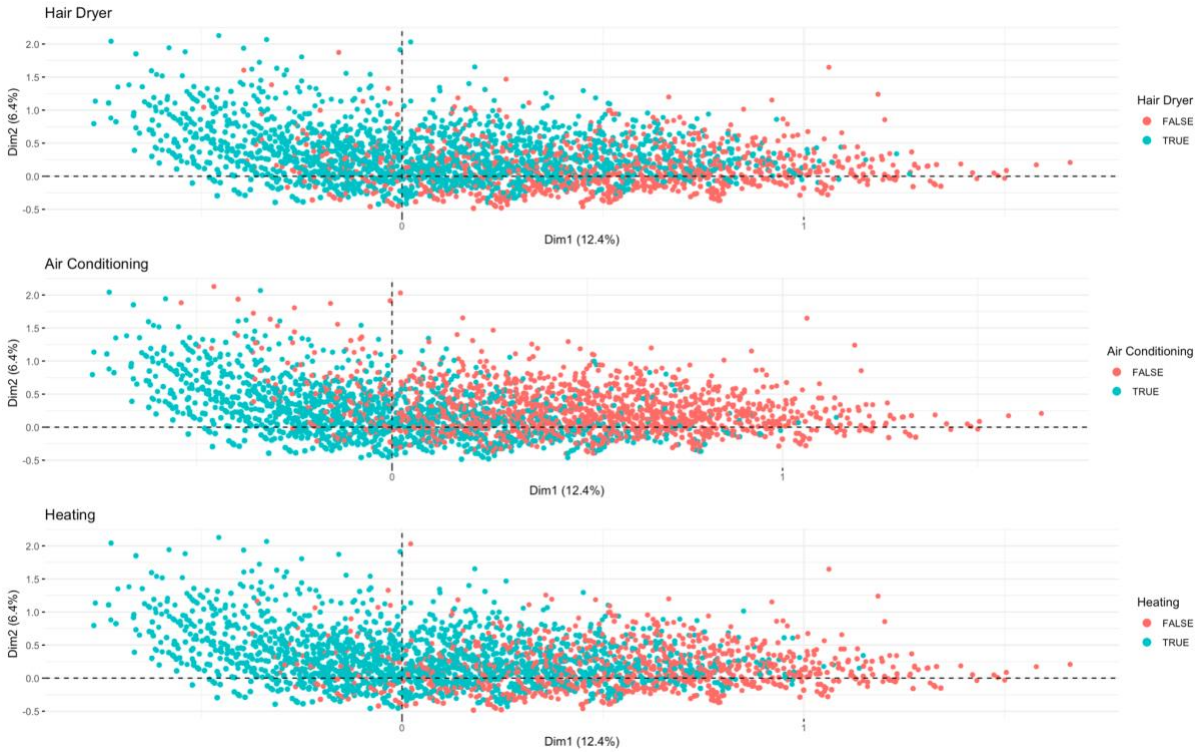
Room type dimensions 1 and 2



Phone, email and work email verifications

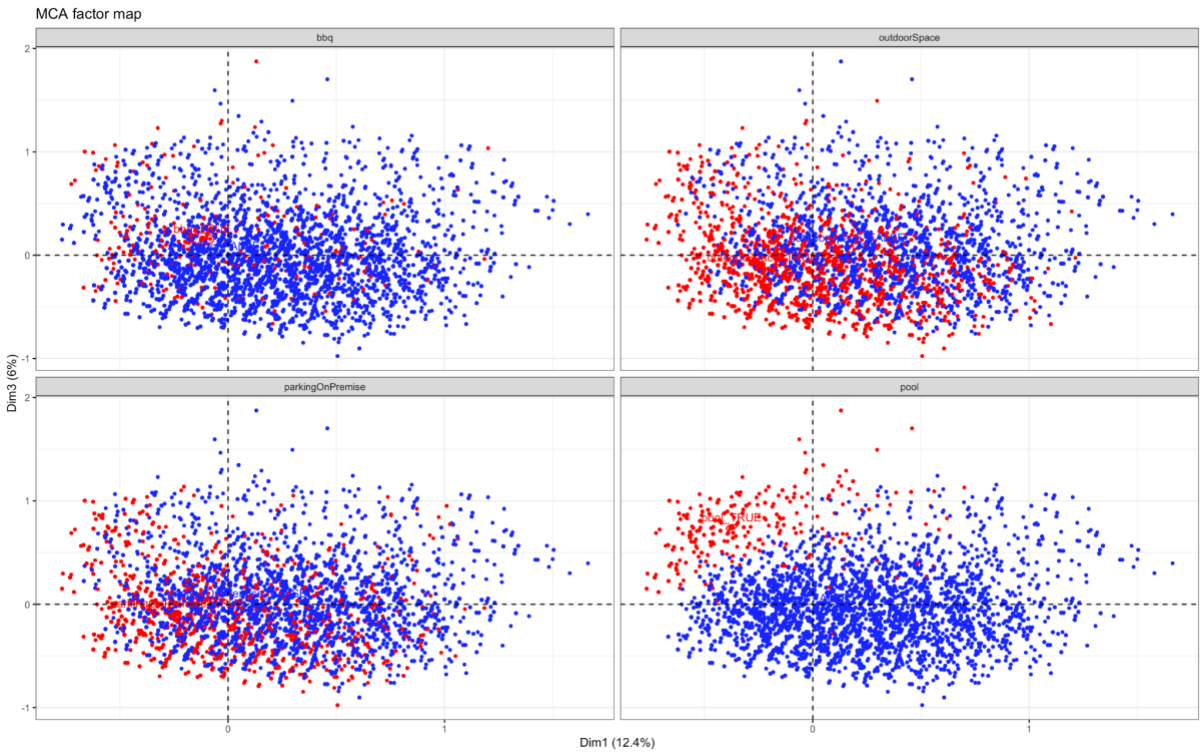


Long term stays and host greets dimensions 1 and 2



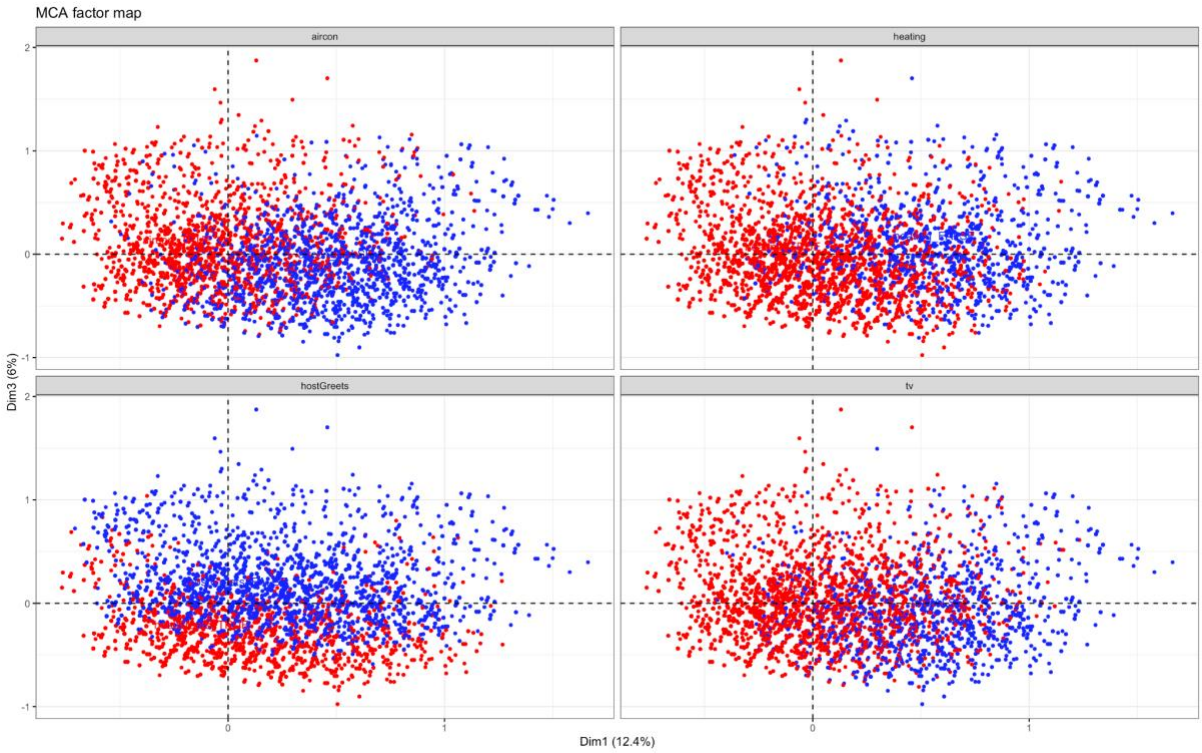
Hair dryer, air conditioning and heating dimensions 1 and 2

Barbecue, outdoor space, parking on premise and pool in dimensions 1 and 2 (red=True, blue=False)

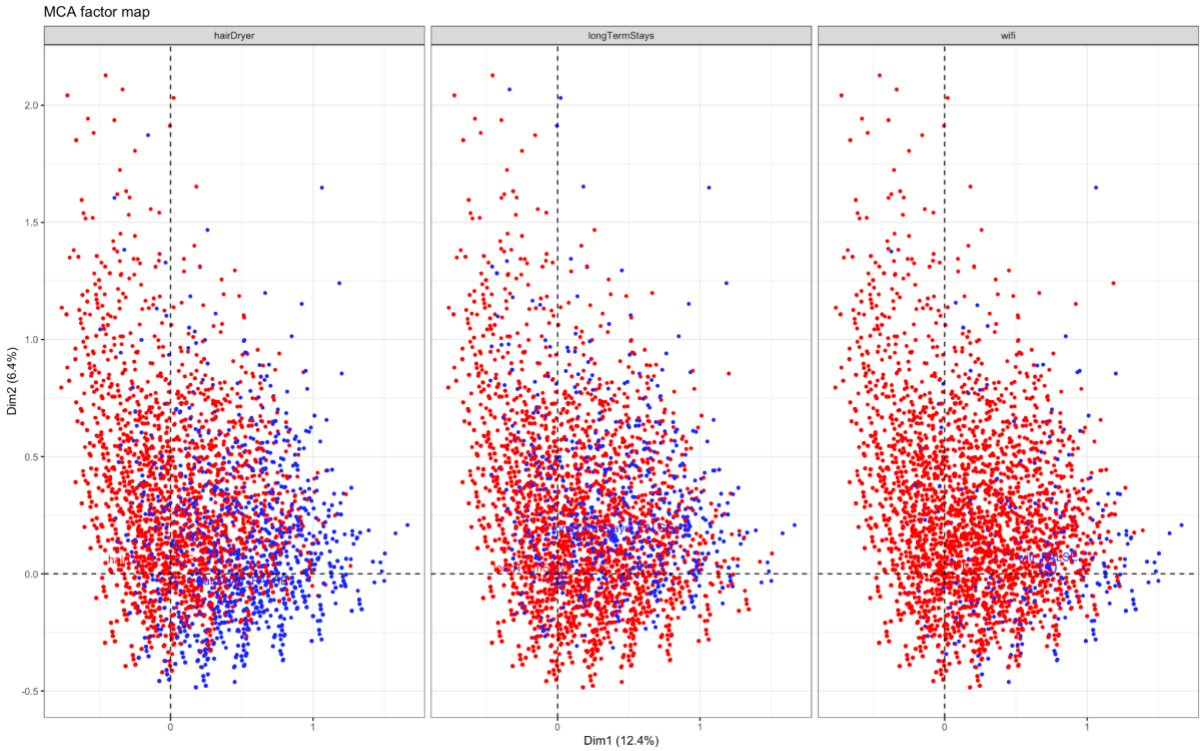


Barbecue, outdoor space, parking on premise and pool in dimensions 1 and 3 (red=True, blue=False)

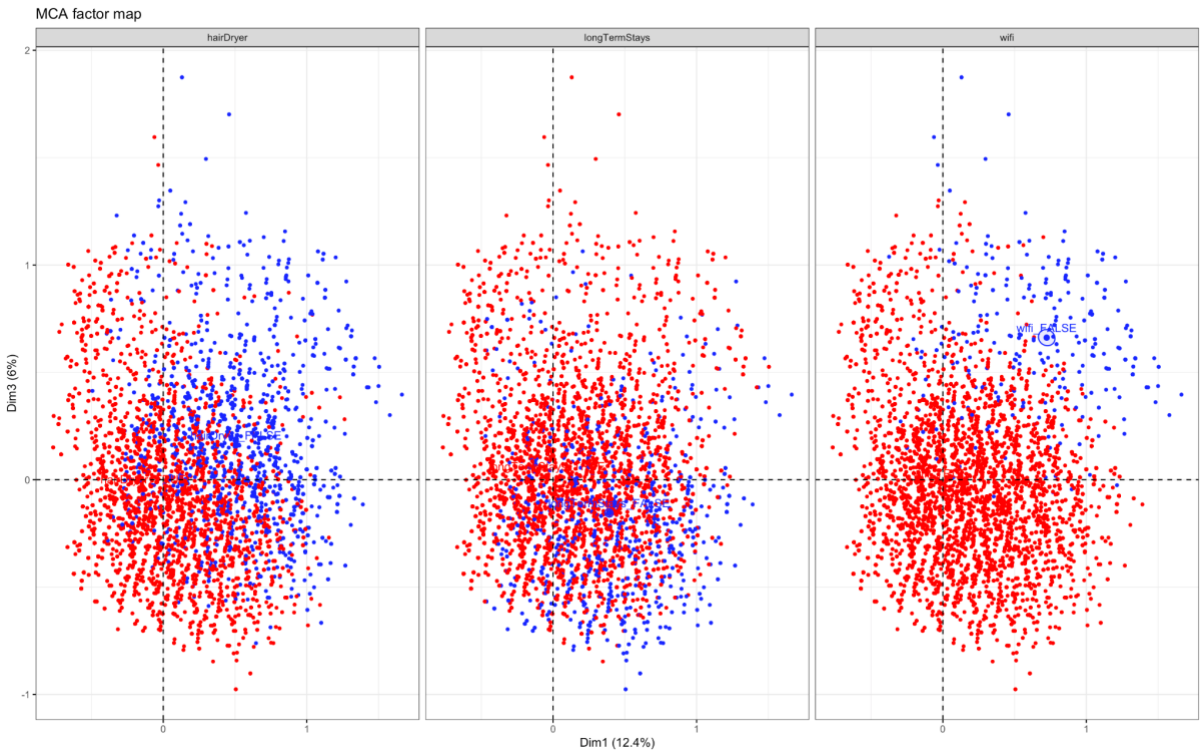
Air conditioning, heating, host greets and tv in dimensions 1 and 2 (red=True, blue=False)



Air conditioning, heating, host greets and tv in dimensions 1 and 3 (red=True, blue=False)



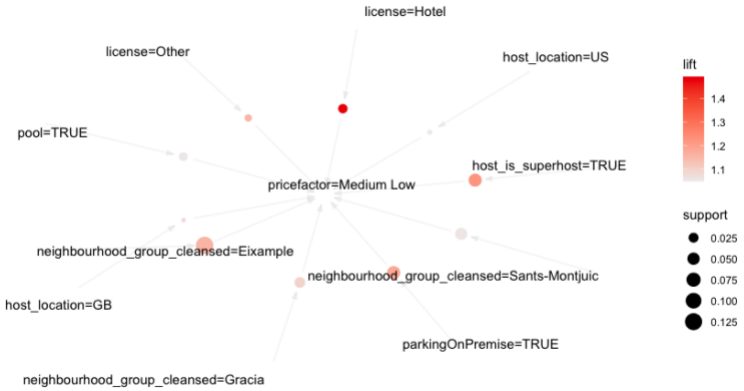
Hair dryer, long term stays, wifi in dimensions 1 and 2 (red=True, blue=False)



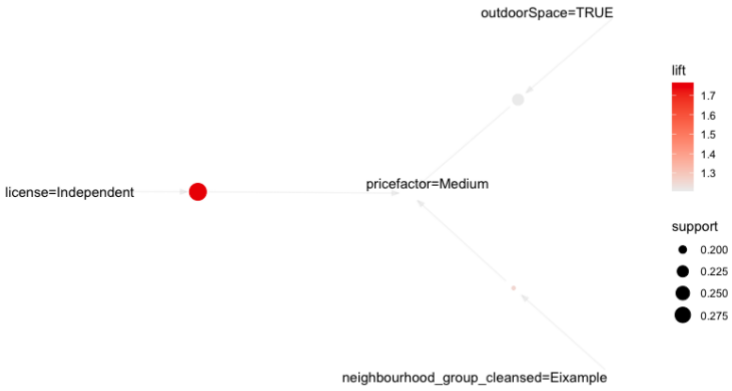
Hair dryer, long term stays, wifi in dimensions 1 and 3 (red=True, blue=False)

ASSOCIATION MAPS

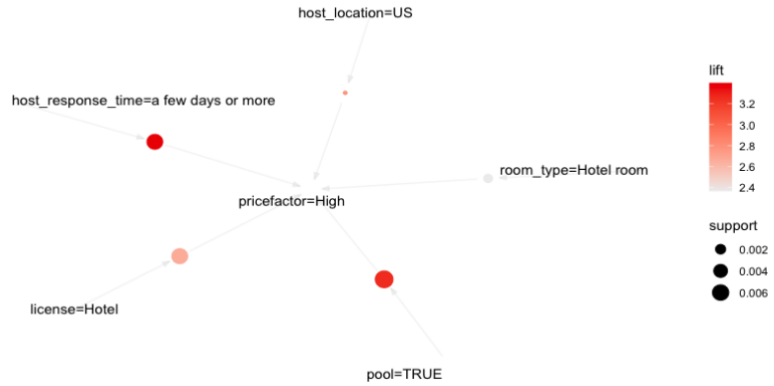
Association rules for Medium Low price with length 2



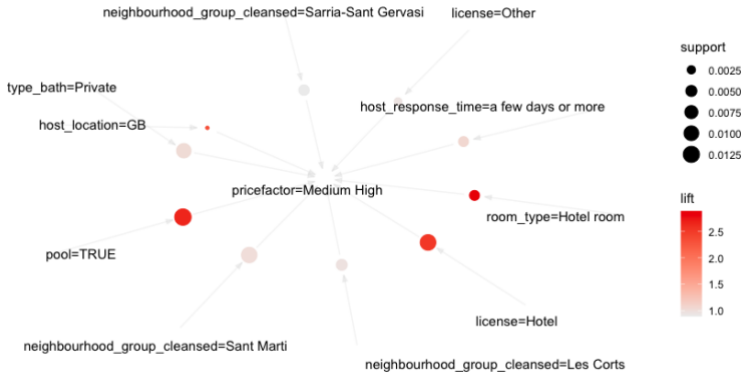
Association Rules for price factor Medium (originally though as Medium before we decided to split it in 2)



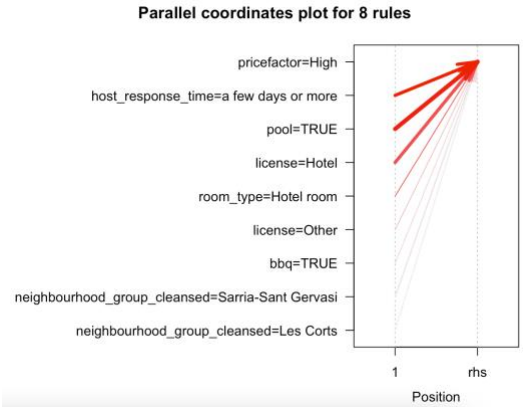
Association Rules for price High



AR for Price Medium High



Parallel Coordinates Plot for 8 rules



Rules of length 5